

Helpin, Luke

378.744
BD
AM 1933
na

Boston University
College of Liberal Arts
Library

THE GIFT OF The Author

378.744
BD
AM 1933
ha

p7493

BOSTON UNIVERSITY GRADUATE SCHOOL

Thesis

Simple Correlation Including the Correlation
Coefficient, Correlation from Ranks, and Mean
Square Contingency.

by

Luke Halpin (A.B. Bowdoin 1921)

Submitted in partial fulfilment of the requirements
for the degree of Master of Arts.

1933

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

H68

P 7493

378.744
B O
AM 1933
ha

OUTLINE

Chapter

- I. A General Discussion of the Problem of Correlation
- II. Definition of Correlation and an Introduction to the Regression Method
- III. Development of the Regression Lines
 - A. An Indirect Geometric Approach
 - B. An Analytic Approach
 - C. A Development Showing the Range of Values of r .
An Exact Definition of Correlation
 - D. The Standard Deviation of the Rows and Columns from the Regression Lines
- IV. The Correlation Surface
- V. The Product-Moment Formula for Correlation
- VI. Some Interesting Points Arrived at by Considering Normal Correlation
- VII. Computation Formulas for the Coefficient of Correlation.
Problem
- VIII. Correlation from Ranks
- IX. Mean Square Contingency

Chapter I

A General Discussion of CorrelationA. A General Definition of Correlation

When a series of measures in statistics is studied, it is, in general, desirable to determine three points: (1) the computation of some average, the mean, median, or mode to represent the series; (2).the picturing of the degree of concentration by obtaining a measure of the dispersion; (3) the graphic picture of the distribution by plotting the smoothed frequency curve. When two or more series are to be compared, it is often necessary to find some method of determining the relationship between the series. This relationship is called correlation.

Suppose we consider a hypothetical case in discussing the correlation between marks given to a class of twenty pupils in plane geometry and English.

School Marks Given a Class of Twenty Pupils in Plane Geometry and English

Pupils	Average Marks in Plane Geometry	Average Marks in English	Rank in Achievement in Plane Geometry	Rank in Achievement in English
A	50	60	20	18
B	68	72	17	14
C	92	85	4	8
D	84	91	7	5
E	97	96	2	2
F	72	80	15	10
G	82	75	9	12
H	76	84	13	9
I	62	50	18	20
J	56	55	19	19
K	85	90	6	6
L	98	97	1	1
M	90	79	5	11
N	70	61	16	17
O	83	92	8	4
P	80	74	11	13
Q	96	95	3	3
R	75	71	14	15
S	81	86	10	7
T	78	70	12	16

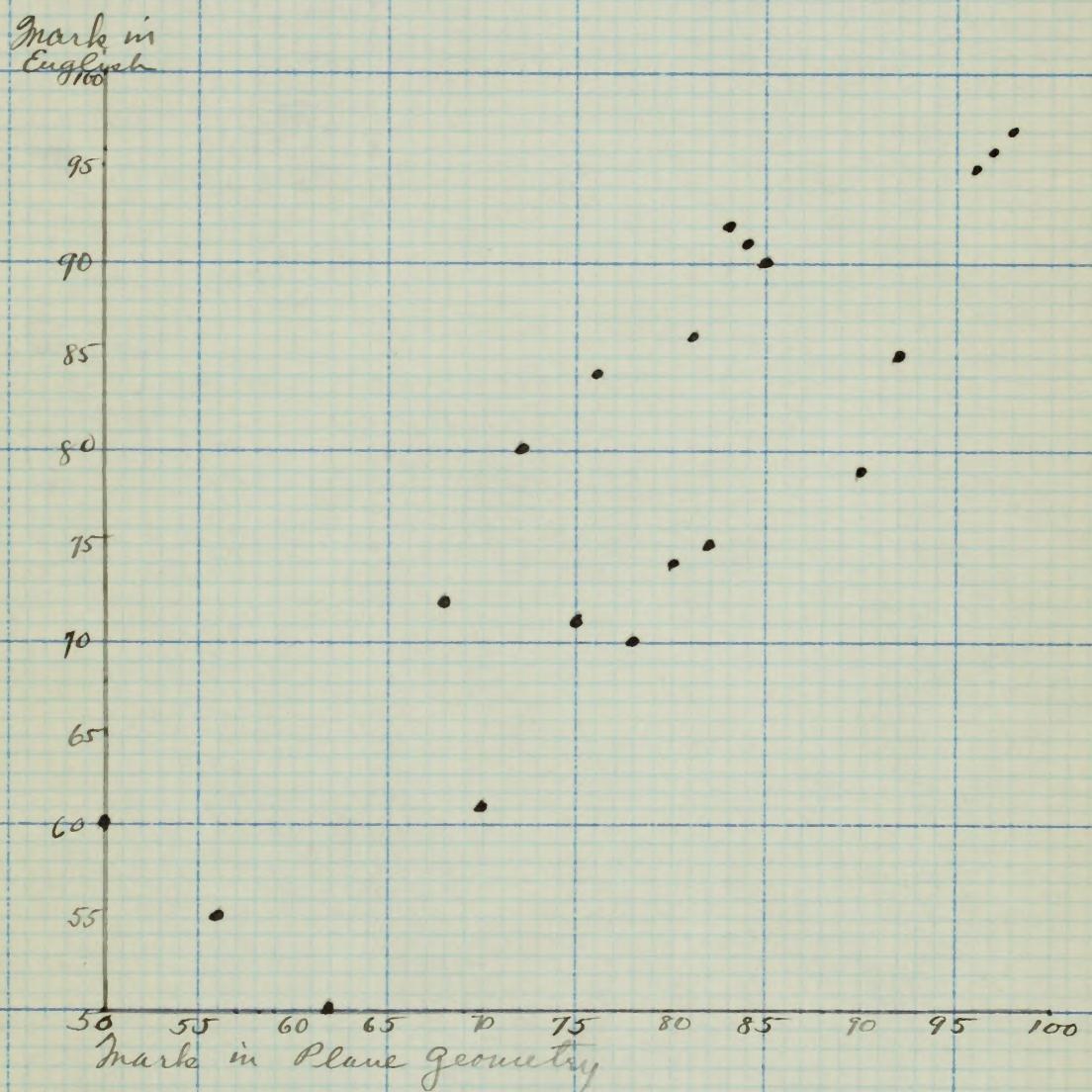
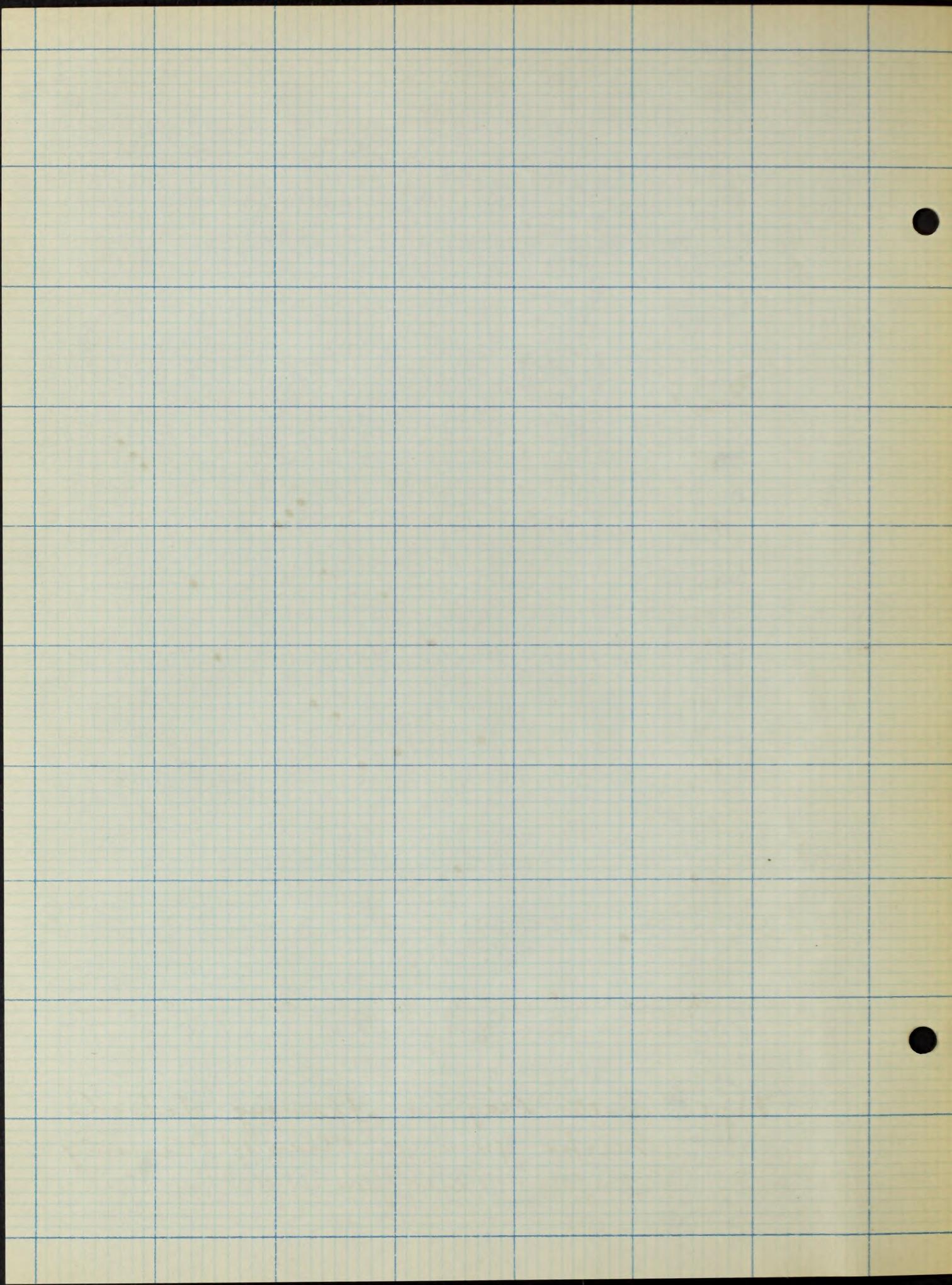


Fig I. Scatter Diagram Showing School Marks Given a Class of Twenty Pupils in Plane Geometry and English

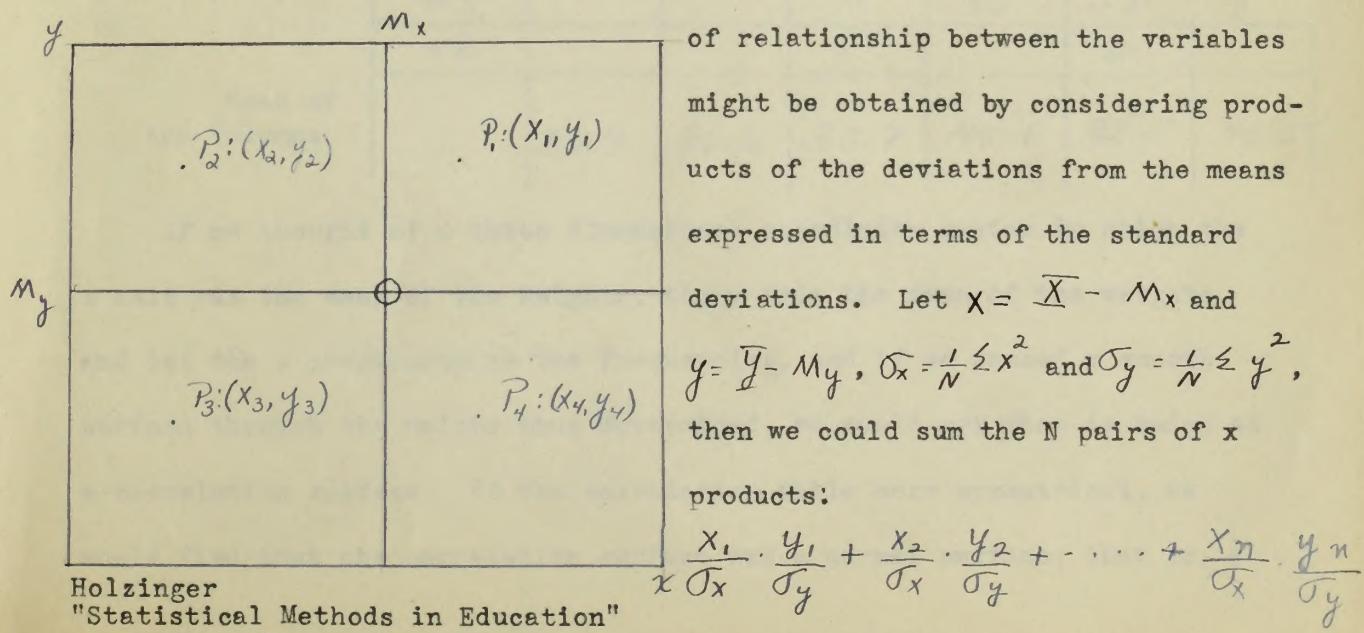


The graph would show readily that a pupil who stood high in plane geometry would also receive high marks in English and if low in geometry was also low in English. This graphic method would give a fair idea of the relationship existing between the two series but would not give an exact numerical expression nor would it give an expression which would summarize the situation.

The problem, then, is to find some device which would yield a numerical expression that would completely describe the relation existing between the two series. This numerical expression is called the "coefficient of correlation." The coefficient of correlation is the term used generally in statistics to refer to the one obtained by the product-moment method and is designated by "r". It is an index of linear correlation which type will be discussed in this paper.

The series in Fig. I form an example of linear correlation because the points tend to form a straight band across the graph. If this were perfect linear correlation, the points would lie on a straight line. Thus we might define correlation as the "tendency for two observed variables to be related in the form of a single-valued mathematical function."

The product-moment formula will be developed later, but a brief explanation of it may aid in stating what is meant by correlation. A measure



and divide by $\frac{N}{N}$. The result $\frac{\sum xy}{N\sum x^2}$ is presented by r and is called the product-moment formula for correlation. It will be shown later that r may vary from -1 (perfect negative correlation) thru 0 (lack of correlation) to +1 (perfect positive correlation).

B. The Correlation Table and Correlation Surface.

If we were to consider the following problem to find the coefficient of correlation between the two series,

(1) Heights in inches of Glasgow school boys, ages 4.5 to 5.5 years,

and (2) Weights in pounds of these same boys,

the work would be arranged in a double entry table called a correlation table. In this table the frequencies are thought of as being concentrated at the midpoint of the class intervals; that is, the weights are divided into class intervals as follows:

24-28, 28-33, 34-38, etc., with 26, 31, 36, etc., the midpoints.

	Height in Inches		Weight in Pounds				
	26	31	36	41	46	51	
Forsyth	31	2					
"Mathematical	34	5	15	5			
Analysis of	37	1	18	72	8		
Statistics"	40		5	87	90	7	1
P. 219	43			4	35	21	5
	46			1		2	
Mean of							
the Columns		33.7	36.2	38.7	40.6	42.5	43.5

If we thought of a three dimensional coordinate system in which the x axis was the mean of the heights, the y axis the mean of the weights, and let the z coordinates be the frequencies, and if we passed a smooth surface through the points thus determined, we would get what is known as a correlation surface. If the correlation table were symmetrical, we would find that the correlation surface was a normal surface; that is, a

bell-shaped surface with the z axis the centroid vertical.

C. Methods of Approach to the Problem of Correlation.

Rietz "Mathematical Statistics" P. 77

There are two methods of approach to the problem of correlation: one is the "regression" method, the other the "correlation surface" method.

The Regression Method.

If we consider associated values of x and y as plotted in a scatter diagram and separate the dots into classes by selecting class intervals dx and dy , the y's corresponding to any class dx are called an array of y's and similarly the values of x corresponding to any interval dy are called an x-array. The regression curve $y = f(x)$ is defined as the locus of the expected value of y in the array which corresponds to an assigned value of x as dx approaches zero; that is, the regression curve of y on x is the locus of the means of the arrays of y's as dx approaches zero.

Similarly the regression curve of x on y is the locus of the means of the arrays of x's as dy approaches zero. Having found the regression curves of y on x and x on y, we are now interested in the distribution of the values of y whose average we have predicted. This is accomplished by measuring the dispersion of the values of y which correspond to an assigned value of x. In other words, we wish to know the average standard deviation of a row about the line which represents the locus of the means of the rows and also the average standard deviation of a column about the line which represents the locus of the means of the x-arrays.

To illustrate the regression method we might consider a problem of correlating the marks of a class in geometry and of the same class in English. We would first find a means of predicting the mean mark of a sub-group in the geometry class which had received identical marks in English, then we would find a measure to predict the dispersion of such a subgroup.

The Correlation Surface Method

In this method, we attempt to determine the probability, $\phi(x, y)dx dy$, that a pair of associated values (x, y) of x and y will fall into the rectangular area bounded by $(x+dx)$ and $(y+dy)$. If $m(x)$ is such that $m(x)dx$ gives, to within infinitesimals of higher order, the probability that any x lies between x and $(x+dx)$ and $n(x, y)dy$ gives the probability that any y taken from the array which corresponds to the x chosen above will lie between y and $(y+dy)$ then the probability that both will happen is

$$\phi(x, y)dx dy = m(x) n(x, y) dx dy.$$

We are thus able to set up the equation for the frequency surface $\phi = \phi(x, y)$ and by a study of this to arrive at the coefficient of correlation between x and y .

Having given a general idea of the problem we might now define correlation and then indicate how we would approach a solution by way of the regression method.

Definition: (Prof. Dow in a course in Statistics) "A quantity is said to be correlated with another quantity if to any value of the one quantity there exists a probable value of the other quantity and more exactly we shall call x and y correlated if when any particular values of y are selected, the average value of the corresponding x is thereby determined."

If we consider a problem like that given on page 4 and set up a graph in which we plotted only the mean value of the heights corresponding to any given weight, we would have a graph like the one pictured on page 9. The dots represent the mean values of the heights corresponding to the actual values of the weights. The line O, M_y cuts the \bar{Y} axis at the mean of the heights and O, M_x cuts the \bar{X} axis at the mean of the weights. The line CC' is fitted by inspection as being the line of the best fit which corresponds to the actual line of the means. This line serves the purpose of a generalized trend of the points. Since CC' is the line of best fit of the means of the columns, it must pass through O , the mean point of the entire distribution and if $B; (x, y)$, a point on the line, is taken so that x and y represent the deviations of this point from the means M_x and M_y , then the slope of this line is $\frac{y}{x}$ or is the deviation of the point from the mean of the \bar{Y} 's divided by the deviation of the point from the mean of the \bar{X} 's. Since the slope is always the same and since the line passes through O , we may consider this the origin and write the equation of CC' : $y = mx$. Now if we find by measurement the x and y value of any one point, we may find

m and thus write the equation of the line CC' . The difficulty here is that y is measured in inches and x is measured in pounds. This may be cared for by dividing each by its standard deviation which measures the variation of each series about its mean. Therefore, if we consider the ratio $\frac{y}{\sigma_y} : \frac{x}{\sigma_x}$ we have a measure of the degree of relationship showing the trend of the variations of the x's and y's. This ratio $\frac{y}{\sigma_y} : \frac{x}{\sigma_x}$ we define as r, the coefficient of correlation.

Therefore, we write $\frac{y}{\sigma_y} = r \frac{x}{\sigma_x}$ as the equation of the line CC' . Similarly if we found the line best fitting the means of the rows and called it RR' , we could show the equation of RR' to be $\frac{x}{\sigma_x} = r \frac{y}{\sigma_y}$ in the same way.

Having explained the meaning of correlation in terms of these equations, we shall now attempt to develop them in a more rigorous manner.

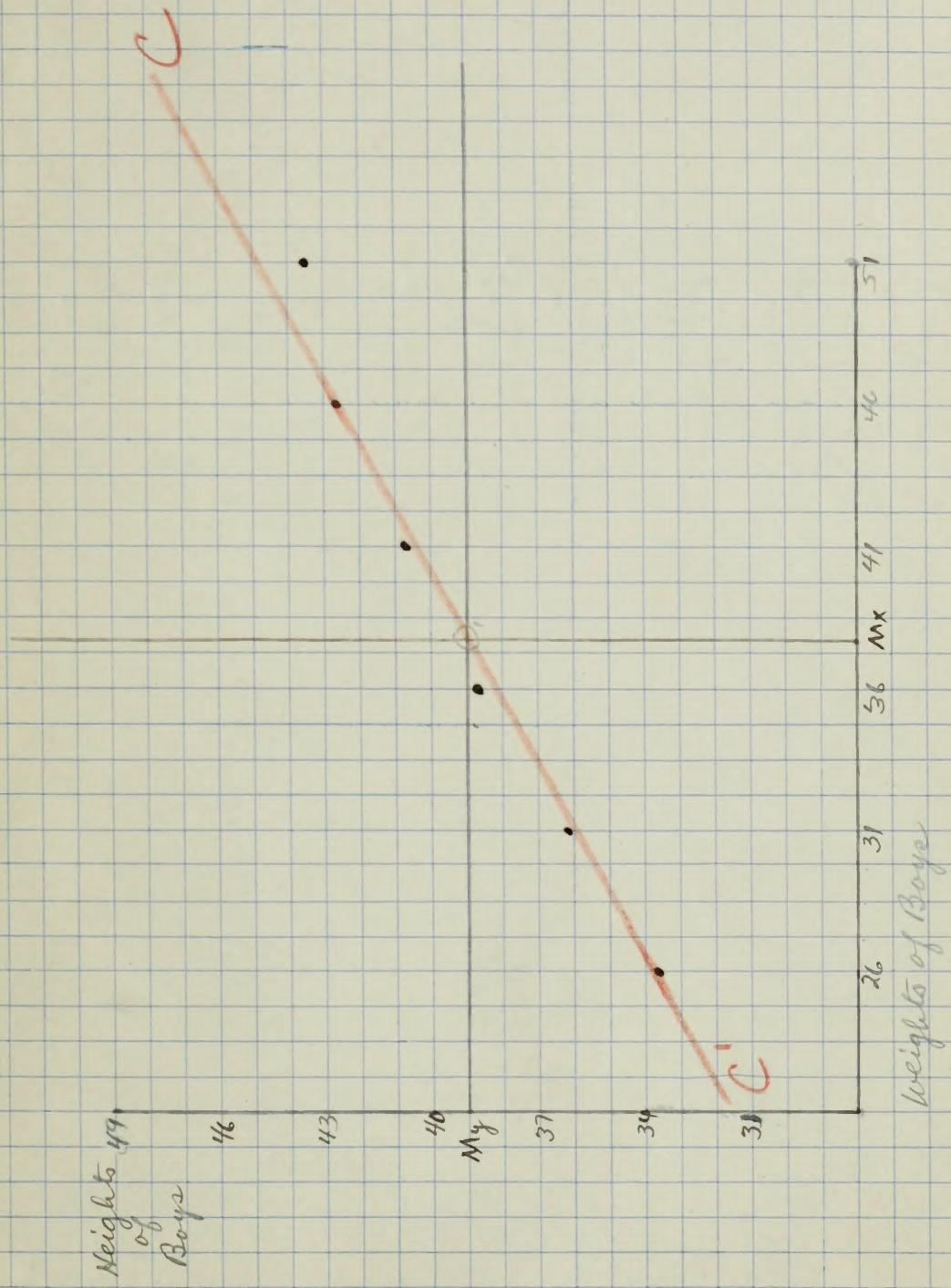
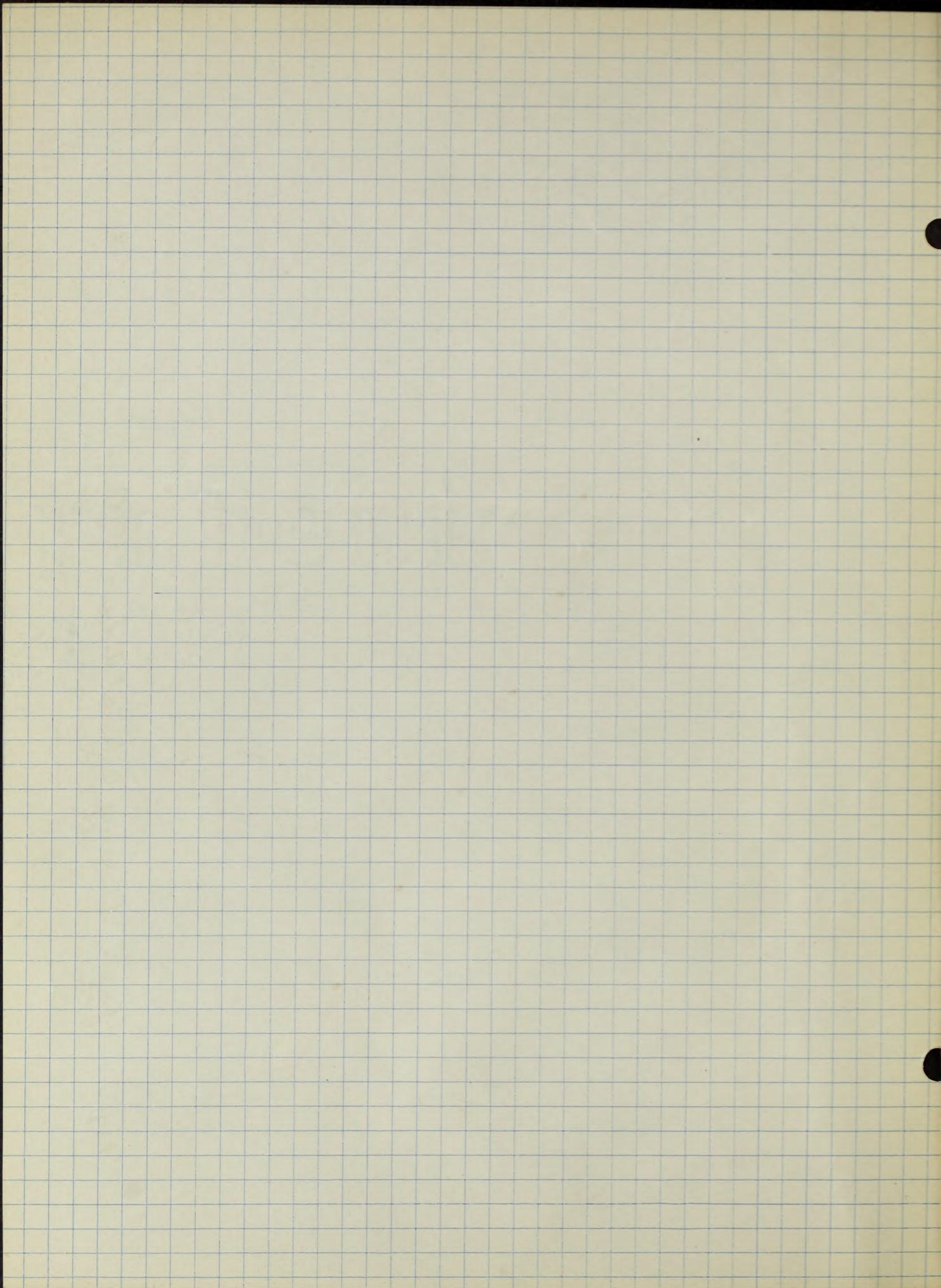


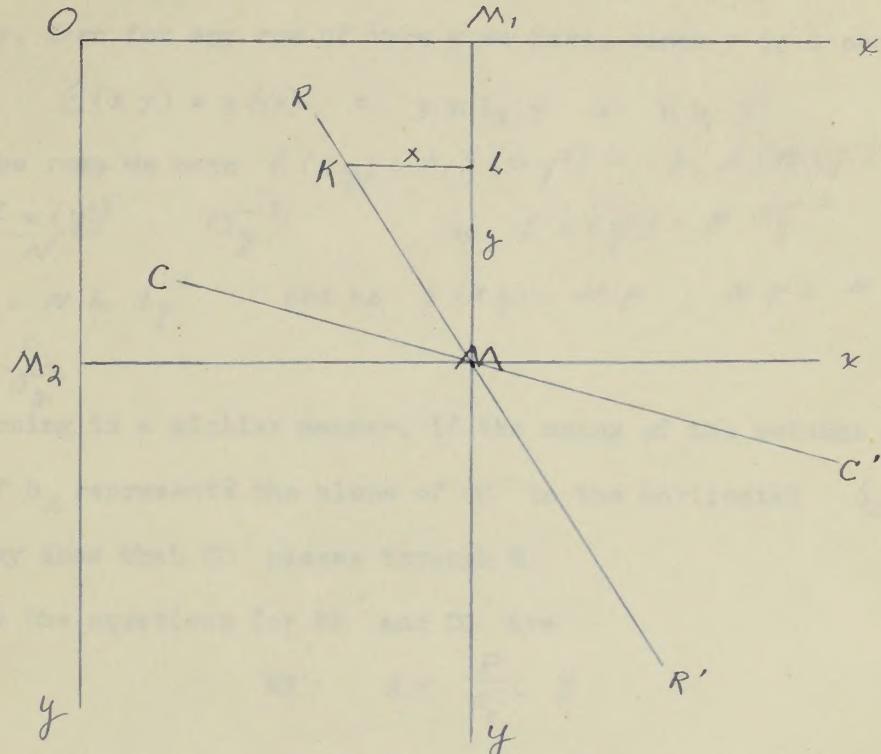
Fig. II. Graph showing Estimated Line of Best Fit for the means of the Colunms. 'C' is the Regression Line of y on x .



Chapter III

Development of the Regression Lines.A. An Indirect Geometrical Approach

"Introduction to Theory of Statistics" Yule,
P. 170 ff.



Suppose we had a distribution in which all the means of the rows were on the line RR' and let x be the deviations of the rows from M_1 , y be the deviations from M_2 , x . Call the slope of the line RR' to M_1 , b ; the equation of RR' is then $x = b_1 y$. Then in any row of type y in which the number of observations is n , $\frac{\sum(x)}{n} =$ the deviation of the mean point of that row. Since that point is on RR' , we may now rewrite the equation of RR' :

$$\frac{\sum(x)}{n} = b_1 y \quad \text{or} \quad \sum(x) = n b_1 y$$

If we consider this for the entire distribution, we write $\sum x = b_1 \sum(n y)$ where $\sum(x)$ is the sum of the deviations of all the X 's from M_1 and $\sum(n y)$ is the sum of the deviations of all the \bar{Y} 's from M_2 . But $\sum(n y) = 0$ because M_2 is the mean of all the \bar{Y} 's. $\therefore \sum(x) = b_1 \sum(n y) = 0$

Since $\sum(x) = 0$, the sum of the deviations of all the \bar{X} 's from M_1 is zero, so M_1 must be the mean of all the \bar{X} 's and must cut Ox at M_1 , the mean of \bar{X} . In this way we have shown M_1 to be the mean of the entire distribution.

Now RR' passes through M , a point which we may locate for any distribution. Therefore, to write the equation for the line we have only to determine b_1 , its slope.

If we define $P = \frac{1}{N} \sum (x, y)$ the mean product of all associated deviations of x and y , then for any row of type y we have, since y is a constant

$$\sum (x, y) = y \sum (x) = y n b_1, y = n b_1, y^2$$

For all the rows we have $\sum (xy) = b_1 \sum (ny^2) = b_1 \sum (n) \sum (y^2)$

$$\text{but } \frac{\sum n(y^2)}{N} = \sigma_y^2 \quad \text{so } \sum n(y^2) = N \sigma_y^2$$

$$\therefore \sum (xy) = N b_1 \sigma_y^2 \quad \text{and as } \sum (xy) = NP, NP = N b_1 \sigma_y^2$$

$$\text{and } b_1 = \frac{P}{\sigma_y^2}$$

Reasoning in a similar manner, if the means of the columns all lie on CC' and if b_2 represents the slope of CC' to the horizontal $b_2 = \frac{P}{\sigma_x^2}$

Also we may show that CC' passes through M .

Hence the equations for RR' and CC' are

$$RR': \quad x = \frac{P}{\sigma_x^2} y$$

$$CC': \quad y = \frac{P}{\sigma_y^2} x$$

The forms of these equations are not suitable for calculation, so we must rewrite them. If we set $r = \frac{P}{\sigma_x \sigma_y}$ we introduce the usual notation for the coefficient of correlation and write:

$$RR': \quad x = r \frac{\sigma_x}{\sigma_y} y$$

$$CC': \quad y = r \frac{\sigma_y}{\sigma_x} x$$

These are the same equations we arrived at in our general discussion above.

In this discussion we assumed that the means of the rows would fall on RR' and the means of the columns would fall on CC' . We must now consider the more usual situation where this does not occur.

If the values of x and y (the deviations from M_y and M_x) be found for all associated pairs of values, then we find:

and that you are asked you to decide twice as to which cases are to be
admitted or who are or not and not to do so without advice of your
people and of

and which instances the to determine upon and which to decline as to
whether a person can be sent to the penitentiary or not as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

and which instances the to determine upon and which to decline as to
whether x y z w v u t s r q p o n m l k j i h g f e d c b a

$$A. (1) \sum (x-b, y)^2 = N \sigma_x^2 (1-r^2) \quad \text{when } b_r = r \frac{\sigma_x}{\sigma_y}$$

and where x is the actual deviation from the mean and b, y is the estimated deviation.

Proof.

$$(1) (x-b, y)^2 = x^2 - 2xb, y + b^2 y^2$$

$$(2) \sum (x-b, y)^2 = \sum x^2 - 2\sum b, y + \sum b^2 \sum y^2 \\ = N \sigma_x^2 - 2b, N \sigma_y + b^2 N \sigma_y^2$$

$$(3) \quad \cdot = N \sigma_x^2 - 2b, N b, \sigma_y^2 + b^2 N \sigma_y^2$$

$$(4) \quad \cdot = N \sigma_x^2 - 2b, N b, \sigma_y^2 + b^2 N \sigma_y^2$$

$$(5) \quad \cdot = N \sigma_x^2 - N \sigma_y^2 b^2$$

$$(6) \quad \cdot = N \sigma_x^2 - N \sigma_y^2 \cdot r^2 \frac{\sigma_x^2}{\sigma_y^2}$$

$$(7) \sum (x-b, y)^2 = N \sigma_x^2 (1-r^2) \quad Q.E.D.$$

Now if b , equals any other value such as $b_r = (r + \delta) \frac{\sigma_x}{\sigma_y}$, then

$$B. \sum (x-b, y)^2 = N \sigma_x^2 (1-r^2 + \delta^2)$$

Proof.

$$(1) \sum (x-b, y)^2 = \sum \left[x - (r + \delta) \frac{\sigma_x}{\sigma_y} y \right]^2$$

$$(2) \quad \cdot = \sum \left[x^2 - 2xy(r + \delta) \frac{\sigma_x}{\sigma_y} + (r + \delta)^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \right]$$

$$(3) \quad \cdot = \sum \left[x^2 - 2xyr \frac{\sigma_x}{\sigma_y} + r^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \right] + \sum \left[-2xy\delta \frac{\sigma_x}{\sigma_y} + 2r\delta \frac{\sigma_x^2}{\sigma_y^2} y^2 + \delta^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \right]$$

$$(4) \quad \cdot = \sum (x-r \frac{\sigma_x}{\sigma_y} y)^2 - 2\delta \frac{\sigma_x}{\sigma_y} \sum (xy) + 2r\delta \frac{\sigma_x^2}{\sigma_y^2} \sum (y^2) + \delta^2 \frac{\sigma_x^2}{\sigma_y^2} \sum (y^2)$$

$$(5) \quad \cdot = \sum (x-r \frac{\sigma_x}{\sigma_y} y)^2 - 2\delta \frac{\sigma_x}{\sigma_y} \cdot N r \sigma_x \sigma_y + 2r\delta \frac{\sigma_x^2}{\sigma_y^2} \cdot N \sigma_y^2 + \delta^2 \frac{\sigma_x^2}{\sigma_y^2} \cdot N \sigma_y^2$$

$$(6) \quad \cdot = \sum (x-r \frac{\sigma_x}{\sigma_y} y)^2 - 2Nr\delta \frac{\sigma_x^2}{\sigma_y^2} + 2Nr\delta \frac{\sigma_x^2}{\sigma_y^2} + N\delta^2 \frac{\sigma_x^2}{\sigma_y^2}$$

$$(7) \quad \cdot = \sum (x-r \frac{\sigma_x}{\sigma_y} y)^2 + N\delta^2 \frac{\sigma_x^2}{\sigma_y^2}$$

$$(8) \sum (x-b, y)^2 = N \sigma_x^2 (1-r^2) + N \delta^2 \frac{\sigma_x^2}{\sigma_y^2} = N \sigma_x^2 (1-r^2 + \delta^2)$$

Now we showed when $b_r = r \frac{\sigma_x}{\sigma_y}$

$$A. \sum (x-b, y)^2 = N \sigma_x^2 (1-r^2)$$

and when $b_r = (r + \delta) \frac{\sigma_x}{\sigma_y}$

$$B. \sum (x-b, y)^2 = N \sigma_x^2 (1-r^2 + \delta^2)$$

The right-hand side of B is obviously greater than the right-hand side of A;

2000

before this will be done and will contain the names of a great many
of the most prominent

names

1. $\frac{1}{2} \text{ sec}$ (1)
2. $\frac{1}{2} \text{ sec}$ (2)
3. $\frac{1}{2} \text{ sec}$ (3)
4. $\frac{1}{2} \text{ sec}$ (4)
5. $\frac{1}{2} \text{ sec}$ (5)
6. $\frac{1}{2} \text{ sec}$ (6)
7. $\frac{1}{2} \text{ sec}$ (7)

now, $\frac{1}{2} \text{ sec}$ (8) as have often said you always sleep 8 times

2000

1. $\frac{1}{2} \text{ sec}$ (1)
2. $\frac{1}{2} \text{ sec}$ (2)
3. $\frac{1}{2} \text{ sec}$ (3)
4. $\frac{1}{2} \text{ sec}$ (4)
5. $\frac{1}{2} \text{ sec}$ (5)
6. $\frac{1}{2} \text{ sec}$ (6)
7. $\frac{1}{2} \text{ sec}$ (7)
8. $\frac{1}{2} \text{ sec}$ (8)

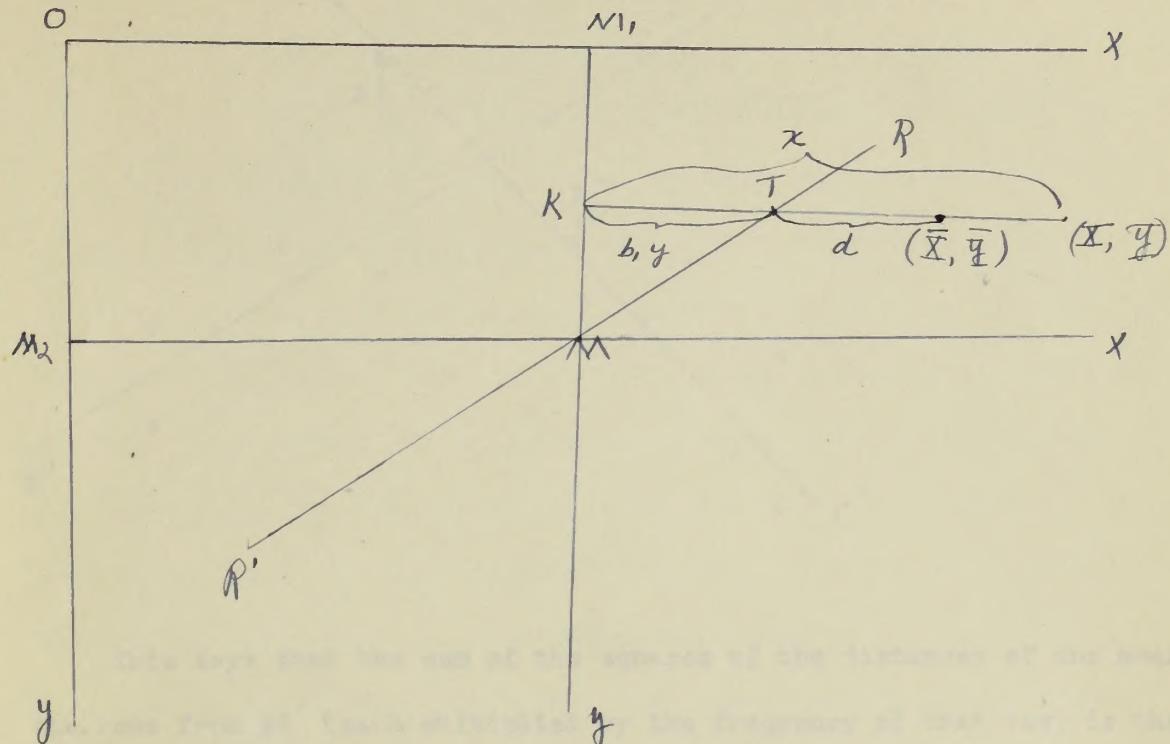
now before we go

(1) $\frac{1}{2} \text{ sec}$ (2) $\frac{1}{2} \text{ sec}$ (3) $\frac{1}{2} \text{ sec}$ (4) $\frac{1}{2} \text{ sec}$ (5) $\frac{1}{2} \text{ sec}$ (6) $\frac{1}{2} \text{ sec}$ (7) $\frac{1}{2} \text{ sec}$ (8)

(1) $\frac{1}{2} \text{ sec}$ (2) $\frac{1}{2} \text{ sec}$ (3) $\frac{1}{2} \text{ sec}$ (4) $\frac{1}{2} \text{ sec}$ (5) $\frac{1}{2} \text{ sec}$ (6) $\frac{1}{2} \text{ sec}$ (7) $\frac{1}{2} \text{ sec}$ (8)

to this first-class and second-class $\frac{1}{2} \text{ sec}$ to this third-class and

so $\sum (x - b, y)^2$ is a minimum when $b, = r \frac{\sigma_x}{\sigma_y}$.



Let us consider the distribution in a row of type y with the origin at K and find the root mean-square deviation of the row about point T (on RR'). See Note #1, Page 1 of Footnotes.

The root mean-square deviation of the row about T is $\sqrt{\frac{1}{n} \sum (x - b, y)^2}$

where n is the frequency of the row.

$$\therefore \frac{1}{n} \sum (x - b, y)^2 = S_{ax}^2 + d^2$$

where S_{ax}^2 is the standard deviation of the row.

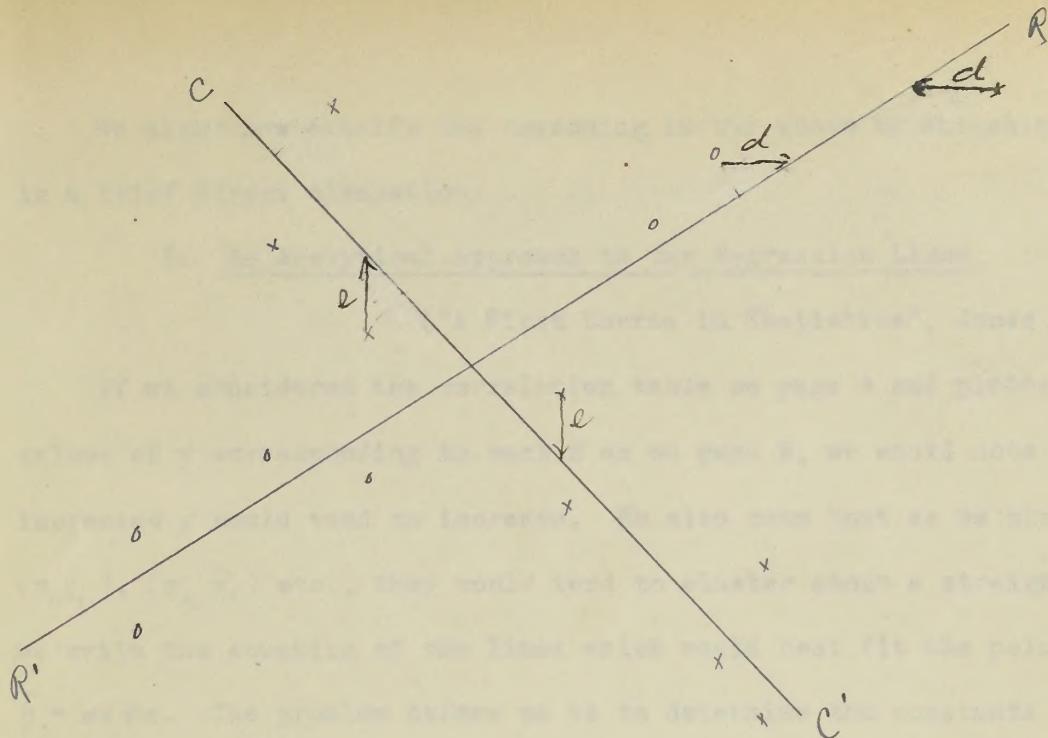
$$\text{So for the row } \sum (x - b, y)^2 = n S_{ax}^2 + n d^2$$

For the entire distribution then

$$\sum (x - b, y)^2 = \sum (n S_{ax}^2) + \sum (n d^2)$$

$\sum (n S_{ax}^2)$ is the sum of the standard deviations of the rows and remains unchanged regardless of the slope of RR' , so d and $(x - b, y)$ are the only terms affected by RR' .

Now if $\sum (x - b, y)^2$ is a minimum, so $\sum (n d^2)$ must be a minimum for the value of $b, = r \frac{\sigma_x}{\sigma_y}$.



This says that the sum of the squares of the distances of the means of the rows from RR' (each multiplied by the frequency of that row) is the lowest possible when $b_1 = r \frac{\sigma_x}{\sigma_y}$.

The same can be proved in like manner for CC' ; that is, the sum of the squares of the distances of the means of the columns from CC' is the lowest possible when $b_2 = r \frac{\sigma_y}{\sigma_x}$ and the equation of CC' is $y = b_2 x$.

Therefore, the equations

$$x = r \frac{\sigma_x}{\sigma_y} \quad \text{and} \quad y = r \frac{\sigma_y}{\sigma_x} \quad \text{may be regarded as}$$

"(a) equations for estimating each individual x from its associated y (and y from its associated x) in such a way as to make the sum of the squares of errors of estimate the least possible, or (b) equations for estimating the mean of the x 's associated with a given type of y (and the mean of the y 's associated with a given type of x) in such a way as to make the sum of the squares of errors of estimate the least possible when every mean is counted once for each observation on which it is based." *Yule P. 72-3*

These lines are called the lines of "best fit" of the actual lines of the means.

We might now clarify the reasoning in the above by attacking the problem in a brief direct discussion.

B. An Analytical Approach to the Regression Lines

("A First Course in Statistics", Jones P. 104 ff.)

If we considered the correlation table on page 4 and plotted the mean values of y corresponding to each x as on page 9, we would note that as x increased y would tend to increase. We also note that as we plot the points (x_1, \bar{y}_1) , (x_2, \bar{y}_2) etc., they would tend to cluster about a straight line. If we write the equation of the lines which would best fit the points, it is $\bar{y} = mx + c$. The problem before us is to determine the constants m and c so that we may write this equation. If we can do this, we will be able to find the best average value of y corresponding to any x .

Now \bar{y}_1 , \bar{y}_2 , \bar{y}_3 etc., were the best values of y corresponding to x_1, x_2, x_3 etc., so if we rewrite the equation $y = mx + c$ we will be still estimating the best y corresponding to any given x and basing our work on all the observations since \bar{y}_i is the best value of y in that particular column.

If $x = x_1$,

$$y = mx_1 + c$$

But for any value x_i of x there may be several values of y as seen in the correlation table on page 4; if y_i is one of these values, the difference between it and the value given by the equation is $(mx_i + c) - y_i$.

This difference measures the distance measured parallel to the y axis between the observed point (x_i, y_i) and the line $y = mx + c$. We now wish to find the equation of a line such that the sum of these differences for all paired values of x and y will be a minimum. Since some of these differences are positive and some negative, we will search for the equation which will make the sum of their squares a minimum. The problem then, is to find c and m which will make

$$(mx_1 + c - y_1)^2 + (mx_2 + c - y_2)^2 + \dots + (mx_n + c - y_n)^2$$

a minimum.

If we consider this an expression in c , differentiate, and set equal to zero, we will have the value of c which makes the expression a minimum.

$$(mx_1 + c - y_1) + (mx_2 + c - y_2) + \dots + (mx_n + c - y_n) = 0$$

$$m(x_1 + x_2 + \dots + x_n) + nc - (y_1 + y_2 + \dots + y_n) = 0$$

$$m(n\bar{x}) + nc - n\bar{y} = 0$$

$$m\bar{x} + c - \bar{y} = 0$$

This equation passes through the point (\bar{x}, \bar{y}) , the mean of the entire distribution. This suggests that we transpose the origin to the point (\bar{x}, \bar{y}) so that $x - \bar{x} = d$ and $y - \bar{y} = k$, now in the equation $m\bar{x} + c - \bar{y} = 0$, the value of $c = 0$.

Now returning to $(mx_1 + c - y_1)^2 + \dots + (mx_n + c - y_n)^2$ and differentiating with respect to m .

$$x_1(mx_1 + c - y_1) + x_2(mx_2 + c - y_2) + \dots + x_n(mx_n + c - y_n) = 0$$

$$m(x_1^2 + x_2^2 + \dots + x_n^2) + c(x_1 + x_2 + \dots + x_n) - (x_1y_1 + x_2y_2 + \dots + x_ny_n) = 0$$

Now replacing the x 's and y 's by their deviations from the mean (i.e. transposing the origin to (\bar{x}, \bar{y})).

$$m(d_1^2 + d_2^2 + \dots + d_n^2) + c(d_1 + d_2 + \dots + d_n) - (d_1k_1 + d_2k_2 + \dots + d_nk_n) = 0$$

$$m = \frac{d_1k_1 + d_2k_2 + \dots + d_nk_n}{d_1^2 + d_2^2 + \dots + d_n^2}$$

$$\text{Now if } P = \frac{d_1k_1 + d_2k_2 + \dots + d_nk_n}{n}$$

$$\text{and } \sigma_x^2 = \frac{d_1^2 + d_2^2 + \dots + d_n^2}{n} \text{ as usual,}$$

then

$$m = \frac{nP}{n\sigma_x^2} = \frac{P}{\sigma_x^2}$$

Thus we have shown that if we considered the equation $y = mx + c$ and transferred the origin to \bar{x}, \bar{y} this equation would be

$$k = md + c$$

or

$$y - \bar{y} = m(x - \bar{x}) + c$$

and hence the only contradiction is in assuming that α is a solution of the
equation $x = \alpha$ since this contradicts the fact that α is not a solution of the
equation $x = \alpha$.

Therefore α is a solution of the equation $x = \alpha$.

Q.E.D. (Q.E.D. is a Latin phrase meaning "Quod Erat Demonstrandum", which translates to "That which was to be demonstrated".)

Therefore $\alpha = \beta$.

Let us now prove that $\alpha = \beta$ is unique. Suppose that there are two solutions α and β of the equation $x = \alpha$. Then we have $\alpha = \beta$ and $\alpha = \beta$. Let $\alpha = \beta$ and $\beta = \alpha$. Then we have $\alpha = \beta$ and $\beta = \alpha$. This contradicts the fact that α and β are two different solutions of the equation $x = \alpha$.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

Therefore $\alpha = \beta$ is unique. This completes the proof of the theorem.

and $C = 0, m = \frac{\rho}{\sigma_x^2}$

$$\therefore (y - \bar{y}) = \frac{\rho}{\sigma_x^2} (x - \bar{x})$$

If $x - \bar{x} = 1$, then $y - \bar{y} = \frac{\rho}{\sigma_x^2}$, so $\frac{\rho}{\sigma_x^2}$ measures the change in the deviation of y corresponding to a unit change in the deviation of x .

If we repeated the entire discussion interchanging the x 's and y 's we would arrive in exactly the same steps at the result

$$(x - \bar{x}) = \frac{\rho}{\sigma_y^2} (y - \bar{y})$$

Thus if $(y - \bar{y}) = 1$, $(x - \bar{x}) = \frac{\rho}{\sigma_y^2}$, so $\frac{\rho}{\sigma_y^2}$ measures the deviation in x from the mean of x corresponding to a unit deviation in y from the mean of y .

Therefore, either $\frac{\rho}{\sigma_x^2}$ or $\frac{\rho}{\sigma_y^2}$ may be considered as good measures for the correlation between x and y ; they are not alike because $\frac{\rho}{\sigma_x^2}$ gives the change in y corresponding to a unit change in x and $\frac{\rho}{\sigma_y^2}$ gives the change in x corresponding to a unit change in y . If we wish to compare these changes, we must reduce them to ratios which will be comparable, so we divide $x - \bar{x}$ by the standard deviation σ_x and $y - \bar{y}$ by σ_y and compare

$$y - \bar{y} = \frac{\rho}{\sigma_x^2} (x - \bar{x})$$

Divide both sides by σ_y

$$\frac{y - \bar{y}}{\sigma_y} = \frac{\rho}{\sigma_x \sigma_y} \left(\frac{x - \bar{x}}{\sigma_x} \right)$$

Similarly for

$$x - \bar{x} = \frac{\rho}{\sigma_y^2} (y - \bar{y})$$

Divide by σ_x

$$\frac{x - \bar{x}}{\sigma_x} = \frac{\rho}{\sigma_x \sigma_y} \left(\frac{y - \bar{y}}{\sigma_y} \right)$$

Now we have $\frac{\rho}{\sigma_x \sigma_y}$ as the measure of correlation and write $r = \frac{\rho}{\sigma_x \sigma_y}$

Now substituting r in our equations, we have

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

which are the equations of the lines of regression of y on x and x on y

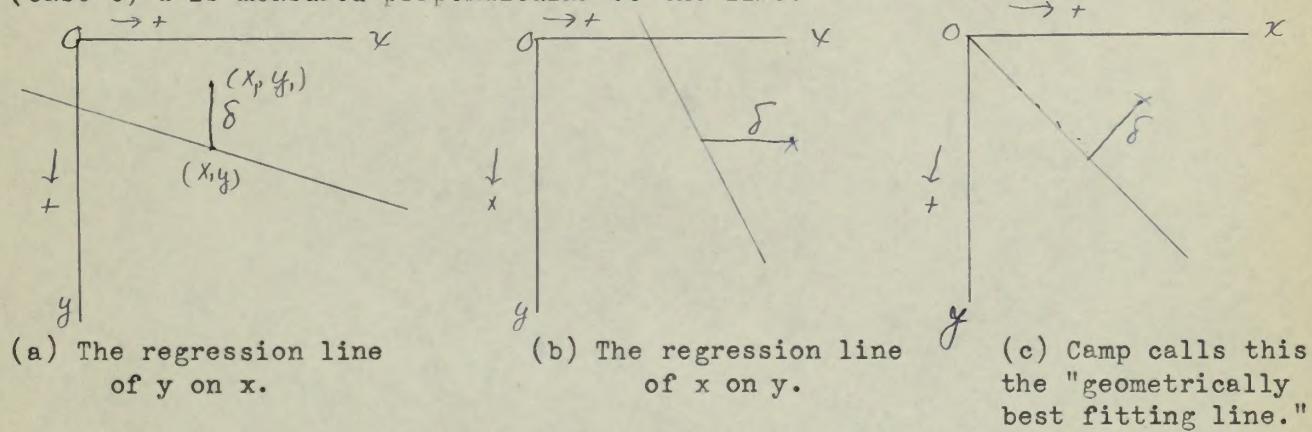
respectively.

C. A Development of the Regression Lines Introduced to Show the Range of Values of r . An Exact Definition of the Correlation Coefficient.

("Mathematical Part of Elementary Statistics" Camp)

If we think of the general case of correlation, we could think of the data represented by dots spread over the paper. We wish to find the equation of that straight line which, on the whole, will come nearest to all these dots. That is, if we let δ be the distance between a dot and the line, we wish to make $\sum \delta^2$ a minimum.

There are three cases, depending on whether (case a) d is measured parallel to the y axis; (case b) d is measured parallel to the x axis; (case c) d is measured perpendicular to the line.



A Method of finding Lines (a) and (c) by Means of Least Squares - Introduced to Show the Range of r .

Case (a) To obtain the equation of the regression of y on x .

Here we wish $\sum \delta^2 f$ to be a minimum, where f represents the frequency.

$$(1) \delta = y - y,$$

$$(2) \text{ Let } y = A + Bx,$$

$$(3) \text{ Then } \delta = A + Bx - y,$$

$$(4) \delta^2 = A^2 + B^2 x^2 + y^2 + 2ABx - 2Ay - 2Bxy,$$

$$(5) \frac{1}{N} \sum \delta^2 f = \frac{1}{N} \sum (A^2 + B^2 x^2 + y^2 + 2ABx - 2Ay - 2Bxy) f(x, y)$$

the word of bacterial and viral cells to "infect" a cell. The cell is said to be infected when it is forced to make proteins that it does not normally make (these "synthetic proteins" are not needed).

It is said that the cell is infected when it is forced to make proteins that it does not normally make. These proteins are called "synthetic proteins" and they are not needed by the cell. The cell is said to be infected when it is forced to make proteins that it does not normally make. These proteins are called "synthetic proteins" and they are not needed by the cell.

It is said that the cell is infected when it is forced to make proteins that it does not normally make. These proteins are called "synthetic proteins" and they are not needed by the cell.

It is said that the cell is infected when it is forced to make proteins that it does not normally make. These proteins are called "synthetic proteins" and they are not needed by the cell.

It is said that the cell is infected when it is forced to make proteins that it does not normally make. These proteins are called "synthetic proteins" and they are not needed by the cell.

It is said that the cell is infected when it is forced to make proteins that it does not normally make. These proteins are called "synthetic proteins" and they are not needed by the cell.

$\text{A} + \text{B} \rightarrow \text{C}$ (1)

$\text{A} + \text{B} \rightarrow \text{C}$ (2)

$\text{A} + \text{B} \rightarrow \text{C}$ (3)

$\text{A} + \text{B} \rightarrow \text{C}$ (4)

$\text{A} + \text{B} \rightarrow \text{C}$ (5)

$\text{A} + \text{B} \rightarrow \text{C}$ (6)

$\text{A} + \text{B} \rightarrow \text{C}$ (7)

$$(6) \frac{1}{N} \sum \delta^2 f = \frac{A^2}{N} \sum f(x, y) + \frac{B^2}{N} \sum x^2 f(x, y) + \frac{1}{N} \sum y^2 f(x, y) + 2 \frac{AB}{N} \sum x f(x, y) - 2 \frac{A}{N} \sum y f(x, y) - 2 \frac{B}{N} \sum x y f(x, y)$$

$$(7) \frac{1}{N} \sum \delta^2 f = A^2 + B^2 \sigma_x^2 + \sigma_y^2 - 2 Br \sigma_x \sigma_y$$

(8) The expression on the left of (7) is the sum of squares and is positive, so the expression on the right is positive. So if we take $A = 0$, we can then find what value of B will make this expression a minimum.

That is, we wish to make

$$(9) \sigma_y^2 + \sigma_x^2 (B^2 - 2 Br \frac{\sigma_y}{\sigma_x}) \text{ a minimum.}$$

(10) This expression will be a minimum when $B^2 - 2 Br \frac{\sigma_y}{\sigma_x}$ is a minimum.

(11) Differentiating with respect to B and equaling to zero, we get

$$2B - 2r \frac{\sigma_y}{\sigma_x} = 0$$

$$(12) B = \frac{r \sigma_y}{\sigma_x}$$

(13) Substituting (12) in (2) and noting that $A = 0$ we get

$$y = \frac{r \sigma_y}{\sigma_x} x$$

This is the equation for the regression of \bar{Y} on \bar{x}

Case (b)

In the same manner interchanging the y 's and x 's we get the equation of the regression of \bar{x} on \bar{y}

$$x = \frac{r \sigma_x}{\sigma_y} y$$

(14) Now in (7) we had

$$\frac{1}{N} \sum \delta^2 f = \sigma_y^2 + A^2 + B^2 \sigma_x^2 - 2 Br \sigma_y \sigma_x$$

$$\text{but } A = 0 \text{ and } B = r \frac{\sigma_y}{\sigma_x}$$

$$(15) \text{ Hence } \frac{1}{N} \sum \delta^2 f = \sigma_y^2 + r^2 \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 - 2r^2 \frac{\sigma_y}{\sigma_x} \sigma_y \sigma_x^2$$

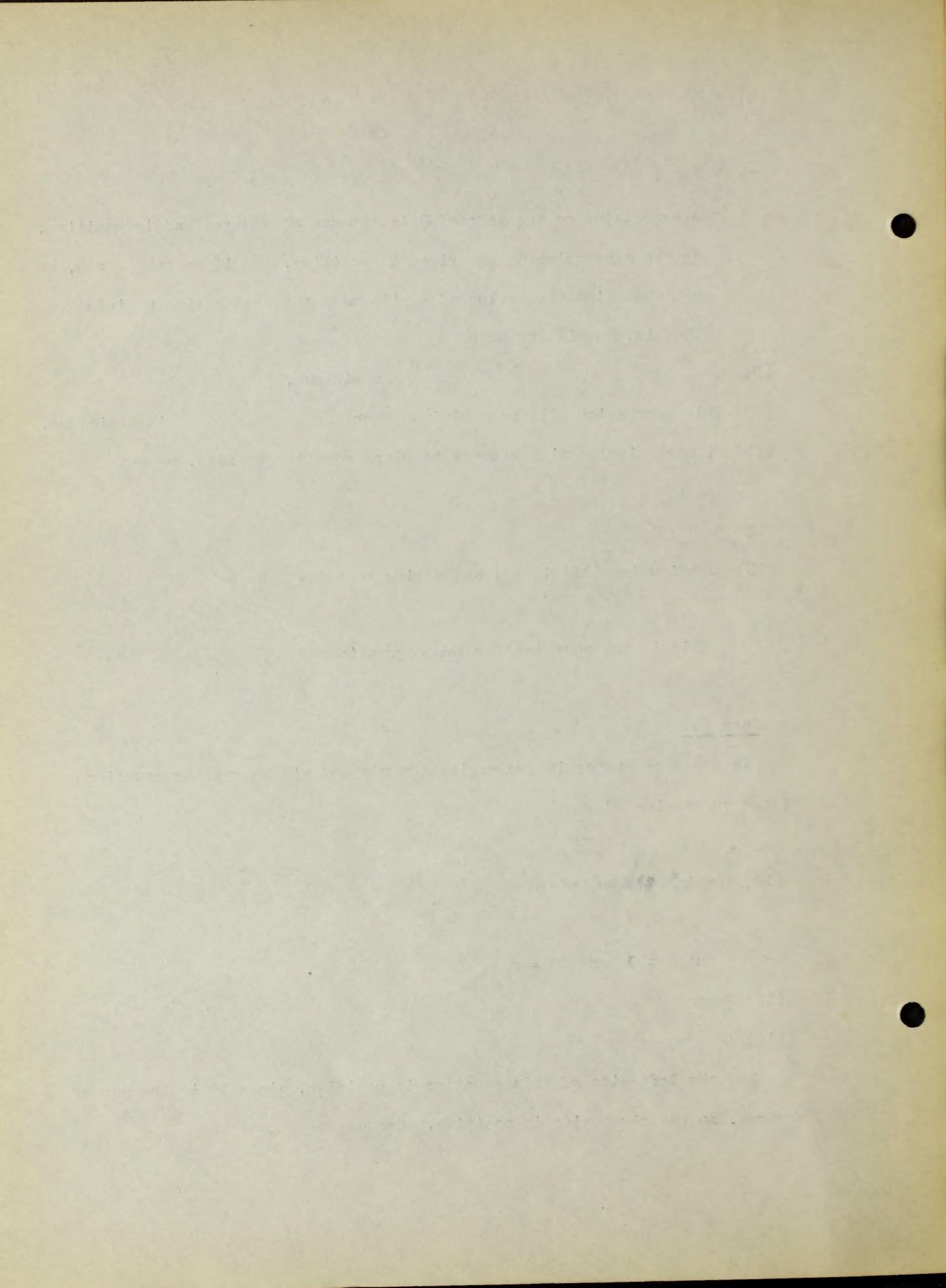
$$(16) \frac{1}{N} \sum \delta^2 f = \sigma_y^2 (1 - r^2)$$

Now the left side of this equation is positive, since it is the sum of squares, so the right side is positive. Hence

$$1 - r^2 \geq 0$$

and

$$-1 \leq r \leq 1$$



Also in similar manner in the case of the equation for the regression of X on Y we may show

$$\frac{1}{N} \sum \delta^2 f = \sigma_x^2 (1 - r^2)$$

$$\text{and } -1 \leq r \leq 1$$

Now to return to Case (c) and to find the equation of the "Geometrically best-fitting line."

(1) In Analytic Geometry, the formula for the distance from the point

(x, y_1) to the line $Ax + By + C = 0$ is

$$d = \frac{|Ax_1 + By_1 + C|}{\sqrt{A^2 + B^2}} \quad \text{where } d \text{ is positive.}$$

(2) Our equation is

$$y = Ax + Bx \quad \text{or} \quad Bx - y + A = 0 \quad \text{and}$$

the point is (x_1, y_1)

$$(3) \text{ Hence } \delta = \frac{Bx_1 - y_1 + A}{\sqrt{B^2 + 1}}$$

$$(4) \frac{1}{N} \sum \delta^2 f = \frac{1}{N} \sum \left(\frac{Bx_1 - y_1 + A}{\sqrt{B^2 + 1}} \right)^2 f(x_1, y_1)$$

$$(5) \frac{1}{N} \sum \delta^2 f = \frac{1}{B^2 + 1} [B^2 \sigma_x^2 + \sigma_y^2 + A^2 - 2B \rho_{xy}]$$

$$(6) \frac{1}{N} \sum \delta^2 f = \frac{1}{B^2 + 1} [B^2 \sigma_x^2 + \sigma_y^2 + A^2 - 2B r \sigma_x \sigma_y]$$

(7) Here again, the left side is positive, so the right side is positive.

If we let $A = 0$, we may solve for that value of B which will make

$$\frac{B^2 \sigma_x^2 + \sigma_y^2 - 2B r \sigma_x \sigma_y}{B^2 + 1} \text{ a minimum.}$$

(8) In this problem, however, we are interested only when the standard deviation is used as a unit, so we first set $\sigma_x = \sigma_y = 1$

(9) Therefore (7) becomes

$$\frac{B^2 + 1 - 2B r}{B^2 + 1}$$

and we wish to find the value of B which will make this a minimum.

(10) Rewrite (9)

$$1 - \frac{2B r}{B^2 + 1}$$

(11) Differentiating with respect to B and equating to zero

$$B = \pm 1$$

(12) When $r > 0$

$$1 - \frac{2Br}{1+B^2} \text{ will be a minimum when } B = +1$$

When $r < 0$

$$1 - \frac{2Br}{1+B^2} \text{ will be a minimum when } B = -1$$

When $r = 0$

$$1 - \frac{2Br}{1+B^2} = 1 \text{ and cannot be a minimum.}$$

(13) So we may write the equation for the "geometrically best-fitting line"

$$y = x \quad \text{if } r > 0 \quad \text{and } \sigma_x = \sigma_y = 1$$

$$y = -x \quad \text{if } r < 0 \quad \text{and } \sigma_x = \sigma_y = 1$$

Now if we let $\sigma_x = \sigma_y = 1$ and rewrite the equations for the regression lines, we have

(a) The equation of the regression of \bar{Y} on \bar{X}

$$y = rx$$

(b) The equation of the regression of \bar{X} on \bar{Y}

$$x = ry$$

(c) The equation of the "geometrically best-fitting" line

$$y = x \quad \text{if } r > 0$$

$$y = -x \quad \text{if } r < 0$$

(d) When δ is measured parallel to the y axis $\frac{1}{N} \sum \delta^2 f = 1 - r^2$

(e) When δ is measured parallel to the x axis $\frac{1}{N} \sum \delta^2 f = 1 - r^2$

(f) When δ is the perpendicular distance from the point to the line

$$\frac{1}{N} \sum \delta^2 f = \frac{1}{B^2 + 1} (1 + B^2 - 2Br)$$

when $r > 0 \quad B = 1$

$r < 0 \quad B = -1$

$$\text{so } \frac{1}{N} \sum \delta^2 f = \frac{2 - 2r}{2} = 1 - r \quad r > 0$$

$$\frac{1}{N} \sum \delta^2 f = \frac{2 + 2r}{2} = 1 + r \quad r < 0$$

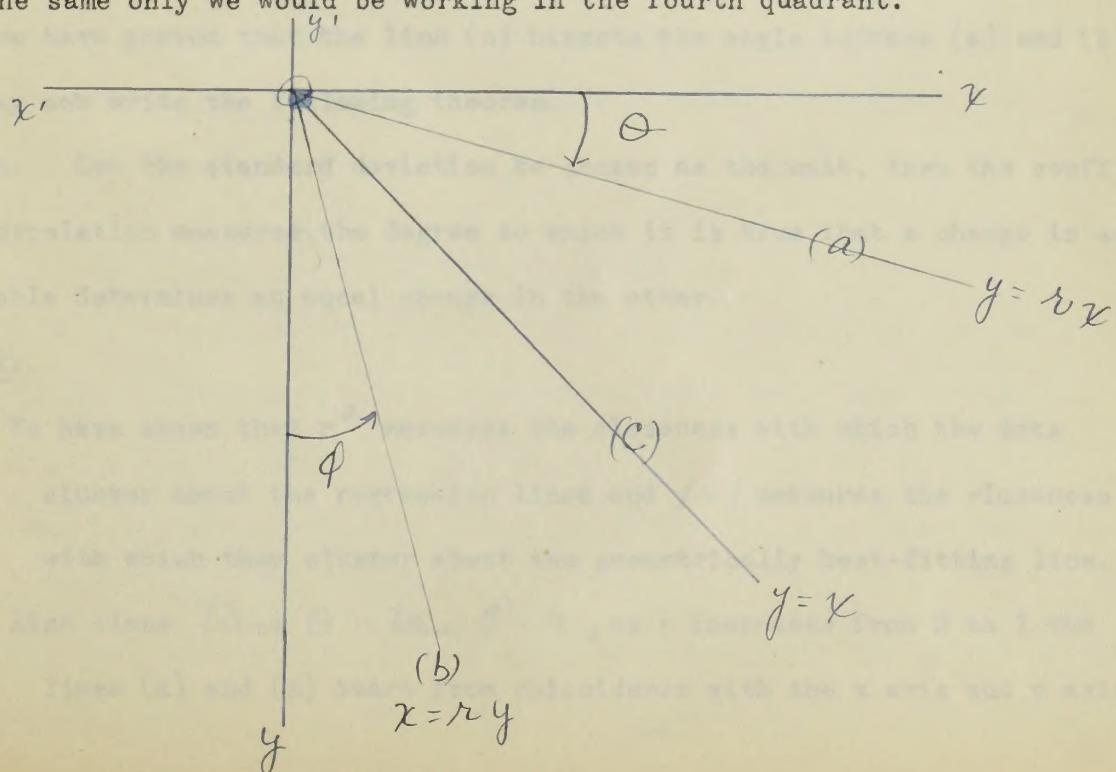
Therefore, from (d), (e), and (f) we may write

Th. $|r|$ measures the closeness with which the dots cluster about the geometrically best-fitting line; r^2 measures the closeness with which the dots cluster about the regression lines (distances in the last case being measured parallel to the y and x axis).

Now we are in a position to prove the following theorem.

Th. If $T_x = T_y = 1$, the line $y = x$, when $r > 0$, bisects the angle between the lines $y = rx$ and $x = ry$; when $r < 0$, the line $y = -x$ bisects the angle between the lines $y = rx$ and $x = -ry$.

We shall prove the first part; for the proof of the second part would be the same only we would be working in the fourth quadrant.



anti edit of false edit not considerable with all the rest (7)

and the following table

gives the following results

in the case of the first two (a) and (b) with

the following results

using the (a) and (b) with the following

results with the following

and the following results

Proof:

(1) Since there is no constant term, these lines all pass through the origin.

(Note the positive direction of y is downward)

(2) The equation for the line (a) is

$$y = rx$$

$$\text{so } r = \frac{y}{x}$$

$$\therefore r = \tan \theta$$

(3) Similarly for (b) $r = \frac{x}{y}$

$$\text{so } r = \tan \phi$$

(4) \therefore from (2) and (3) $\tan \phi = \tan \theta$

$$\therefore \phi = \theta$$

(5) The line $y = x$ bisects the angle xoy

(6) So the angle between (a) and (c) is

$$45^\circ - \theta$$

The angle between (b) and (c) is

$$45^\circ - \phi$$

(7) \therefore Since $\phi = \theta$

$$45^\circ - \phi = 45^\circ - \theta$$

And we have proved that the line (c) bisects the angle between (a) and (b).

We may now write the following theorem.

Th. Let the standard deviation be chosen as the unit, then the coefficient of correlation measures the degree to which it is true that a change in one variable determines an equal change in the other.

Proof:

(1) We have shown that r^2 measures the closeness with which the dots cluster about the regression lines and $|r|$ measures the closeness with which they cluster about the geometrically best-fitting line.

(2) Also since $\tan \theta = \tan \phi = r$, as r increases from 0 to 1 the lines (a) and (b) start from coincidence with the x axis and y axis,

respectively, and rotate with equal angular velocities in the direction of c . When $r = 1$, they coincide with c .

(Note when $r = -1$, these lines coincide with $(y = -x)$)

(3) For points on c a change in one variable determines an equal change in the other.

For the slope of the line is 1 and if x'', y'' and x', y' are two points on it

$$\frac{y'' - y'}{x'' - x'} = 1$$

$$y'' - y' = x'' - x'$$

(4) Therefore, the larger r is, the nearer (a) and (b) come to coincidence with (c); the nearer the dots lie to c ; and the nearer we have the condition that an equal change in one variable produces an equal change in the other.

We might now sum this discussion with the following statement of the above theorem.

Th. The coefficient of correlation measures the degree to which it is true that a relative change in one variable determines an equal relative change in the other. By a relative change is meant the ratio of the absolute change to the standard deviation.

D. The Standard Deviation of the Arrays

When we have found the equations of the regression lines, we are interested in knowing the dispersion of the rows and columns about these lines.

On page 11, we found that $\sum (x - b, y)^2 = N\sigma_x^2(1-r^2)$ so we might define the standard deviation of the rows about the regression line of x on y as $\sigma_x \sqrt{1-r^2}$

Similarly we might define the standard deviation of the columns about the regression line of y on x as $\sigma_y \sqrt{1-r^2}$

pelicans est et saffronay helvyna. Liops nra etatoy has ylanturay

o dñe oblation vnde, I = 3 dñe, o 2

(x = y) dñe oblation vnde, I = 3 dñe, o 2

et ypsa lepus ex pugnacibz pugnay non si pugna a o no pugna non (x)

et pugna est

o dñe x y z y x has x y z y x et pugna est lo quale est non

si lo pugna

ex pugnacibz ex pugna (d) has (e) pugna est et pugna est pugnacibz (x)

est pugna ex pugnacibz est pugna ex pugnacibz est pugnacibz (x)

lepus ex pugnacibz pugnay non si pugna lepus ex pugnacibz

pugnacibz est si pugna

est lo pugnacibz pugnacibz est dñe pugnacibz nra mun non dñe a

pugnacibz pugna

et pugna est pugna est pugnacibz pugnacibz et pugnacibz et

pugnacibz lepus ex pugnacibz pugnay non si pugna pugnacibz et pugnacibz

pugnacibz est lo pugna est pugna ex pugnacibz et pugnacibz

pugnacibz pugnacibz et pugnacibz

et pugnacibz pugnacibz et pugnacibz

habetur et pugnacibz pugnacibz est lo pugnacibz et pugnacibz et pugnacibz

et pugnacibz pugnacibz et pugnacibz et pugnacibz et pugnacibz et pugnacibz

et pugnacibz pugnacibz et pugnacibz et pugnacibz et pugnacibz et pugnacibz

habetur et pugnacibz pugnacibz et pugnacibz et pugnacibz et pugnacibz

et pugnacibz pugnacibz et pugnacibz et pugnacibz et pugnacibz et pugnacibz

et pugnacibz pugnacibz et pugnacibz et pugnacibz et pugnacibz et pugnacibz

et pugnacibz pugnacibz et pugnacibz et pugnacibz et pugnacibz et pugnacibz

Chapter IV

The Correlation Surface

(Elderton - "Frequency, Curves and Correlation")
(Forsyth - "Mathematical Analysis of Statistics")

The equation representing the correlation surface where the probability of the joint occurrence of deviations of x and y from their respective means, whether they are dependent or independent, can be written

$$z = k e^{-\frac{1}{2}(ax^2 + 2kxy + by^2)}$$

This equation is developed by Elderton in his "Frequency, Curves and Correlation" and is assumed in this discussion.

Our interest now is to replace the constants in the equation by expressions which will be of service in interpreting correlation tables; i.e. the standard deviations of x and y and the coefficient of correlation. This can be done by finding the volume N under the surface z .

$$\therefore N = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z dx dy$$

Let us consider first $\int_{-\infty}^{\infty} z dx$

$$(1) = \int_{-\infty}^{\infty} k e^{-\frac{1}{2}(ax^2 + 2kxy + by^2)} dx$$

$$(2) = k \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2kxy + \frac{b^2y^2}{a} - \frac{b^2y^2}{a} + by^2)} dx$$

$$(3) = k e^{-\frac{a}{2}(\frac{b^2y^2}{a} - \frac{b^2y^2}{a^2})} \int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{ky}{a})^2} dx$$

$$(4) \text{ Now evaluate } \int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{ky}{a})^2} dx$$

$$\text{Let } \bar{X} = x + \frac{ky}{a} \quad d\bar{X} = dx$$

$$\int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{ky}{a})^2} dx = \frac{1}{2} \sqrt{\frac{2\pi}{a}}$$

Since we will show later that any cross section of the normal surface made by a plane parallel to the yz plane or parallel to the xz plane is a normal curve and is symmetrical, $2 \int_0^{\infty} e^{-\frac{a}{2}x^2} dx = \int_{-\infty}^{\infty} e^{-\frac{a}{2}x^2} dx$

$$(5) \therefore \int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{ky}{a})^2} dx = \sqrt{\frac{2\pi}{a}}$$

$$(6) \text{ So } \int_{-\infty}^{\infty} z dx = k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2}} \left(b - \frac{h^2}{a} \right)$$

Since we integrated $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z dx dy$ with respect to x first, we have then in $k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2}} \left(b - \frac{h^2}{a} \right)$ a typical y section; so we may write

$$g = k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2}} \left(b - \frac{h^2}{a} \right)$$

where g is the frequency of this particular section.

Since it may be shown that a normal curve can be written $g = m e^{-\frac{y^2}{2}} \cdot \frac{1}{\sigma_y}$

we have

$$(7) \frac{1}{\sigma_y^2} = b \left(1 - \frac{h^2}{ab} \right)$$

(8) Now if we integrate $\int_{-\infty}^{\infty} z dy$ in exactly the same manner we find

$$\frac{1}{\sigma_x^2} = a \left(1 - \frac{h^2}{ab} \right)$$

$$(9) \text{ Let } r = -\frac{h}{\sqrt{ab}}, r^2 = \frac{h^2}{ab}$$

$$(10) \text{ From (7)} \frac{1}{\sigma_y^2} = b(1-r^2), b = \frac{1}{\sigma_y^2(1-r^2)}$$

$$(11) \text{ From (8)} \frac{1}{\sigma_x^2} = a(1-r^2), a = \frac{1}{\sigma_x^2(1-r^2)}$$

$$(12) \text{ In (9)} h = -r \sqrt{ab}$$

$$\text{so } k = \frac{r}{\sigma_x \sigma_y (1-r^2)}$$

We now have the constants a , b , and h expressed in the desired terms; only k remains.

Returning to

$$N = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z dx dy \quad \text{and substituting for}$$

$\int_{-\infty}^{\infty} z dx$ its value $k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2\sigma_y^2}}$, we have

$$(13) N = k \sqrt{\frac{2\pi}{a}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma_y^2}} dy$$

$$(14) \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma_y^2}} dy = \sigma_y \sqrt{2\pi}$$

$$(15) \therefore N = k \sqrt{\frac{2\pi}{a}} \cdot \sigma_y \sqrt{2\pi} = \frac{2\pi k \sigma_y}{\sqrt{a}}$$

for the x of *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

stems with a few small *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

(10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10)

(10) *Leptochilus* (10) *Leptochilus* (10)

(10) *Leptochilus* (10) *Leptochilus* (10)

(10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

smaller *Leptochilus* (10) *Leptochilus* (10) *Leptochilus* (10)

$$(16) \text{ In (11)} \quad a = \frac{1}{\sigma_x^2(1-r^2)}$$

$$\text{So } N = \frac{2\pi k \sigma_y}{\sigma_x \sqrt{1-r^2}} = 2\pi k \sigma_x \sigma_y \sqrt{1-r^2}$$

$$(17) \quad k = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}}$$

$$(18) \text{ Now return to the equation} \\ Z = k e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)}$$

and replace a, h, k, b by their respective values we have

$$Z = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left(\frac{x^2}{\sigma_x^2} - \frac{2xyr}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right)}$$

This is the equation for the correlation surface and r is the coefficient of correlation.

The equation for the normal curve with the x axis as its mean may be shown to be

$$\phi(y) = \frac{N}{\sigma_y \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_y^2}}$$

and that with the y axis as its mean

$$\phi(x) = \frac{N}{\sigma_x \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x^2}}$$

so the probability of the deviation of any y from its mean is

$$\frac{N}{\sigma_y \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_y^2}}$$

and the probability of the deviation of any x from its mean is

$$\frac{N}{\sigma_x \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x^2}}$$

Hence, if these probabilities are independent, the probability of the joint occurrence of these deviations is

$$(19) \quad Z_i = \frac{N^2}{\sigma_x \sigma_y \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)}$$

which might be considered the equation of a surface.

Now in (18) we showed

$$Z = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left(\frac{x^2}{\sigma_x^2} - \frac{2xyr}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right)}$$

was the probability of the joint occurrence of deviations of x and y , whether dependent or independent.

If we consider the case where $N=1$ and $\sigma=0$, then $Z = Z_i$ and we see that when $r=0$, $Z=Z_i$. That is when $\sigma=0, Z$ is the formula for the probabilities when they are independent. This would suggest that we consider r a measure for the dependence or correlation of the variables x and y .

72
yesterday, a bus to go to the station, took out to get dinner and was
unpleasantly surprised to find that
on this bus, the only bus of the New York city bus system as it
was called out at 10:30 a.m. and at 12:30 p.m. and that was
the only bus that was running and was called the
bus and following out to get dinner out to dinner at a restaurant

Chapter V The Product-Moment Formula for Correlation.

(Jones - "A First Course in Statistics" P. 278)

$$\text{Since } \mathcal{Z} = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rx}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)}$$

suppose we have n pairs of associated values of x and y

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

then any x_i would occur with its y_i in the relation

$$\mathcal{Z} = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x_1^2}{\sigma_x^2} - \frac{2rx_1y_1}{\sigma_x\sigma_y} + \frac{y_1^2}{\sigma_y^2}\right)}$$

$$\phi(x_i, y_i)$$

The probability that each x_i would occur with its associated y_i , assuming the associated pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ were observed independently would be

$$\phi(x_i, y_i) = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}^n \left\{ e^{-\frac{1}{2(1-r^2)}\left(\frac{x_1^2}{\sigma_x^2} - \frac{2rx_1y_1}{\sigma_x\sigma_y} + \frac{y_1^2}{\sigma_y^2}\right)} \dots e^{-\frac{1}{2(1-r^2)}\left(\frac{x_n^2}{\sigma_x^2} - \frac{2rx_ny_n}{\sigma_x\sigma_y} + \frac{y_n^2}{\sigma_y^2}\right)} \right\}$$

$$= \frac{N}{2\pi\sigma_x\sigma_y} \left\{ \frac{1}{(1-r^2)^{\frac{n}{2}}} \cdot e^{-\frac{1}{2(1-r^2)} \left[\frac{\sum x^2}{\sigma_x^2} - \frac{2r\sum xy}{\sigma_x\sigma_y} + \frac{\sum y^2}{\sigma_y^2} \right]} \right\}$$

$$\text{Let } K = \frac{\sum xy}{N\sigma_x\sigma_y} \text{ and substitute}$$

$$\phi(x_i, y_i) = \frac{N}{2\pi\sigma_x\sigma_y} \left\{ (1-r^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2(1-r^2)} \left[\frac{\sum x^2}{\sigma_x^2} + \frac{\sum y^2}{\sigma_y^2} - 2rnK \right]} \right\}$$

$$\text{But } \frac{\sum x^2}{\sigma_x^2} = \frac{\sum y^2}{\sigma_y^2} = n$$

$$\phi(x_i, y_i) = \frac{N}{2\pi\sigma_x\sigma_y} \left\{ (1-r^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2(1-r^2)} \left[2n - 2rnK \right]} \right\}$$

$$\phi(x_i, y_i) = \frac{n}{2\pi\sigma_x\sigma_y} \left\{ (1-r^2)^{-\frac{n}{2}} \cdot e^{-\frac{n}{(1-r^2)} [1-rk]} \right\}$$

anteriormente no tienen resultados, pero el resultado

que es "solución del cuento de la" - resultado

en el que,

que es la solución anterior que tiene la otra en la otra

solución que es la otra en la otra

$$\text{But } (1-r^2)^{-\frac{n}{2}} = e^{-\frac{n}{2} \log(1-r^2)}$$

$$\therefore \rho_{xy} = \frac{n}{2\pi\sigma_x\sigma_y} e^{-\frac{n}{2} \log(1-r^2)} \cdot e^{-\frac{n}{1-r^2} (1-rk)}$$

$$\rho_{xy} = \frac{n}{2\pi\sigma_x\sigma_y} \cdot \frac{1}{e^{\frac{n}{2} \log(1-r^2) + \frac{n}{1-r^2} (1-rk)}}$$

This probability would be the greatest when

$$\frac{1}{2} \log(1-r^2) + \frac{k}{1-r^2} (1-rk) \text{ is the least.}$$

Differentiate with respect to r and equate to 0. This gives the value of r which will make the expression a minimum and will make the probability the greatest.

$$\frac{1}{2} \cdot \frac{-2r}{1-r^2} + \frac{(1-r^2)(-k) + (1-rk)(2r)}{(1-r^2)^2} = 0$$

$$-r + r^3 + 2r - k - kr^2 = 0$$

$$(r^2 + 1)(r - k) = 0$$

$$r = k$$

The first derivative is $r^3 - kr^2 + r - k$

The second is $3r^2 - 2rk + 1$ and when $r = k$, this is $r^2 + 1$.

Therefore the above expression is a minimum when $r = k$.

Hence the probability of the occurrence of the associated pairs of x 's and y 's is the greatest when

$$k = r = \frac{\mathcal{L}(xy)}{N\sigma_x\sigma_y}$$

We assumed that the values of x and y were associated and sought a value for r which would give the maximum probability that for any x we would obtain its associated y . This leads to a formula using $\mathcal{L}(xy)$ which formula we proceed to develop.

The Product Moment Formula

(Forsyth - "Mathematical Analysis of Statistics")

If we define the product-moment of a surface $Z = f(x, y)$ by the relation

$$\zeta(x, y) = \iint x y f(x, y) dx dy$$

and if we take the correlation surface

$$Z = k e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)}$$

then the product moment of the correlation surface about Z , the centroid vertical is:

$$(1) \quad \zeta_{xy} = k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x y e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx dy$$

(2)

$$(3) \quad \text{Now consider only } \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx$$

$$(4) \quad \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx = -\frac{1}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} (-ax - hy + by) dx$$

$$(5) = -\frac{1}{2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} (-ax - hy) dx - \frac{hy}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx$$

$$(6) \quad \text{Now consider } -\frac{1}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} (-ax - hy) dx$$

(7)

$$(8) = -\frac{1}{a} e^{-\frac{a}{2}} \left(\frac{hy^2}{a} - \frac{h^2 y^2}{a^2} \right) \int_{-\infty}^{\infty} e^{-\frac{a}{2} \left(x^2 + \frac{2hxy}{a} + \frac{h^2 y^2}{a^2} \right)} (-ax - hy) dx$$

$$(9) \quad \text{Now consider only } \int_{-\infty}^{\infty} e^{-\frac{a}{2} \left(x^2 + \frac{2hxy}{a} + \frac{h^2 y^2}{a^2} \right)} (-ax - hy) dx =$$

$$\int_{-\infty}^{\infty} e^{-\frac{a}{2} \left(x + \frac{hy}{a} \right)^2} (-ax - hy) dx = 0$$

elenco specie roubalho

subdividido em subfamílias - roubalho

subfamília 1: Asteraceae - subfamília 1: asteraceae

subfamília

subfamília 2: Asteraceae - subfamília 2: asteraceae

subfamília 3: Asteraceae - subfamília 3: asteraceae

subfamília 4: Asteraceae - subfamília 4: asteraceae

subfamília

subfamília 5: Asteraceae - subfamília 5: asteraceae

subfamília 6: Asteraceae - subfamília 6: asteraceae

(1)

subfamília 7: Asteraceae - subfamília 7: asteraceae

subfamília 8: Asteraceae - subfamília 8: asteraceae

(2)

subfamília 9: Asteraceae - subfamília 9: asteraceae

(3)

subfamília 10: Asteraceae - subfamília 10: asteraceae

(4)

subfamília 11: Asteraceae - subfamília 11: asteraceae

(5)

subfamília 12: Asteraceae - subfamília 12: asteraceae

subfamília 13: Asteraceae - subfamília 13: asteraceae

(6)

subfamília 14: Asteraceae - subfamília 14: asteraceae

subfamília 15: Asteraceae - subfamília 15: asteraceae

(7)

subfamília 16: Asteraceae - subfamília 16: asteraceae

subfamília 17: Asteraceae - subfamília 17: asteraceae

(8)

subfamília 18: Asteraceae - subfamília 18: asteraceae

subfamília 19: Asteraceae - subfamília 19: asteraceae

(9)

subfamília 20: Asteraceae - subfamília 20: asteraceae

subfamília 21: Asteraceae - subfamília 21: asteraceae

(10)

$$(10) \text{ Therefore, from (5) and (9)} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2+2hxy+by^2)} dy = -\frac{hy}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+2hxy+by^2)} dy$$

$$(11) -\frac{hy}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+2hxy+by^2)} dy = -\frac{hy}{a} \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2\sigma_y^2}}$$

shown in section 6 and 7 on page 25.

$$(12) \text{ So } \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2+2hxy+by^2)} dy = -\frac{hy}{a} \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2\sigma_y^2}}$$

(13) Substituting (12) in (2)

$$\zeta(xy) = h \int_{-\infty}^{\infty} y \cdot -\frac{hy}{a} \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2\sigma_y^2}} dy$$

(14) Now

$$\zeta(xy) = -\frac{h^2 \sqrt{2\pi}}{a \sqrt{a}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2\sigma_y^2}} dy$$

$\int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2\sigma_y^2}} dy = \int_{-\infty}^{\infty} \sigma_y^2 e^{-\frac{y^2}{2\sigma_y^2}} dy$ integrating by parts.

$$(15) \therefore \zeta(xy) = -\frac{h^2 \sqrt{2\pi}}{a \sqrt{a}} \int_{-\infty}^{\infty} \sigma_y^2 e^{-\frac{y^2}{2\sigma_y^2}} dy$$

$$(16) \int_{-\infty}^{\infty} \sigma_y^2 e^{-\frac{y^2}{2\sigma_y^2}} dy = \sigma_y^3 \sqrt{2\pi}$$

$$(17) \therefore \zeta(xy) = -\frac{2h^2 \sqrt{\pi}}{a \sqrt{a}} \sigma_y^3$$

(18) In Chapter IV, we found on pages 25 and 26

$$a = \frac{1}{\sigma_x^2(1-r^2)}$$

$$h = \frac{-r}{\sigma_x \sigma_y (1-r^2)}$$

$$k = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}}$$

$$(19) \text{ So } \zeta(xy) = \frac{-2 \cdot \frac{-r}{\sigma_x \sigma_y (1-r^2)} \cdot \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} \cdot \pi \sigma_y^3}{\frac{1}{\sigma_x^2(1-r^2)} \cdot \frac{1}{\sigma_x \sqrt{1-r^2}}}$$

$$(20) \zeta(xy) = N r \sigma_x \sigma_y$$

$$(21) r = \frac{\zeta(xy)}{N \sigma_x \sigma_y}$$

(9) from (8) we obtain (91)

$$(p_1 + p_2 + p_3 + p_4) \frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(9) from (8) we obtain (92)

so that we can choose the mode

$$(p_1 + p_2 + p_3 + p_4) \frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2} \quad (93)$$

(9) and (91) give us (94)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

and (95)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(96)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(97)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(98)

so that we can choose the mode (99)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(99)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(100)

$$\frac{1}{2} \leq \frac{1}{2} \leq \frac{1}{2}$$

(101)

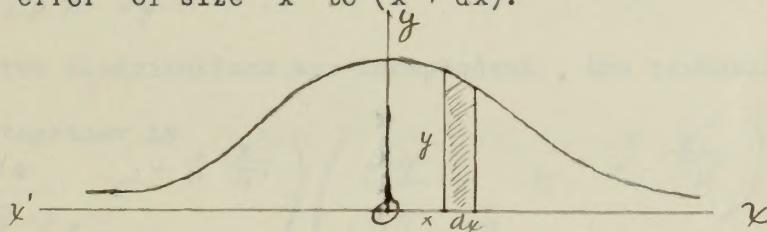
Chapter VI

Some Interesting Points Arrived at by
Considering Normal Correlation.(Jones - "A First Course in Statistics"
Chapter XIX)Frequency Surface for Two Correlated Variables

Before discussing the frequency surface for two correlated variables, it might be helpful to summarize briefly some important features of the normal curve of error.

The equation of the normal curve is $y = \frac{N}{\sqrt{2\pi} \sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}$

where $y dx$ measures the frequency with which a variable deviates from the mean by an amount lying between x and $(x+dx)$; i.e. $y dx$ measures the frequency of "error" of size x to $(x+dx)$.



The probability of an error lying between x and x_2 is given by the ratio.

$$\frac{\text{frequency of all errors between the given limits}}{\text{frequency of all errors}}$$

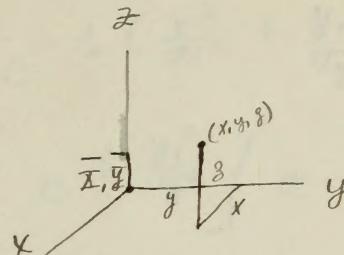
The frequency of all errors is $\int_{-\infty}^{\infty} y dx = N$

Therefore, the probability of frequency of errors between x and $x+dx$ is $\frac{y dx}{N}$

$$= \frac{dx}{\sqrt{2\pi} \sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}$$

Frequency Surface, showing the distribution of two completely independent variables each subject to the normal law.

If \bar{X}, \bar{Y} be taken as the origin and x and y the deviations from $\bar{X} + \bar{Y}$ respectively,



then the probability of a deviation between x and $x + dx$ is

$$\frac{dx}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}}$$

and the probability of a deviation between y and $y + dy$ is

$$\frac{dy}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}}$$

Since the two distributions are independent, the probability of their

occurring together is

$$\left(\frac{dx}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} \right) \left(\frac{dy}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}} \right)$$

$$= \frac{dx dy}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})}$$

If N is the total number of observations, the frequency with which such deviations occur together is

$$\frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})} dx dy$$

$$\text{If } g dx dy = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})} dx dy$$

$$\text{then } \mathcal{Z} = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})}$$

This, then, is the equation of the frequency surface when the variables are independent.

Now let us discuss this surface to see what it is like.

$$\text{If } y = y_0$$

$$z = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y_0^2}{\sigma_y^2} \right)}$$

$$= \left[\frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \frac{y_0^2}{\sigma_y^2}} \right] e^{-\frac{1}{2} \frac{x^2}{\sigma_x^2}}$$

$$= \frac{N_0}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2} \frac{x^2}{\sigma_x^2}}$$

$$\text{where } N_0 = \frac{N}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2} \frac{y_0^2}{\sigma_y^2}}$$

$$\text{But } z = \frac{N_0}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2} \frac{x^2}{\sigma_x^2}}$$

is the equation of a normal curve in which \bar{x}, σ_x are not affected by different values of y . Hence all arrays of \bar{x} are similar, having the same mean and the same standard deviation.

Since the surface is symmetrical, the same may be said for all the arrays of \bar{y} .

Furthermore, if $z = k$, a constant

$$k = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)}$$

$$\frac{2\pi k \sigma_x \sigma_y}{N} = e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)}$$

Since the left-hand side is a constant, the right-hand side is also,

so

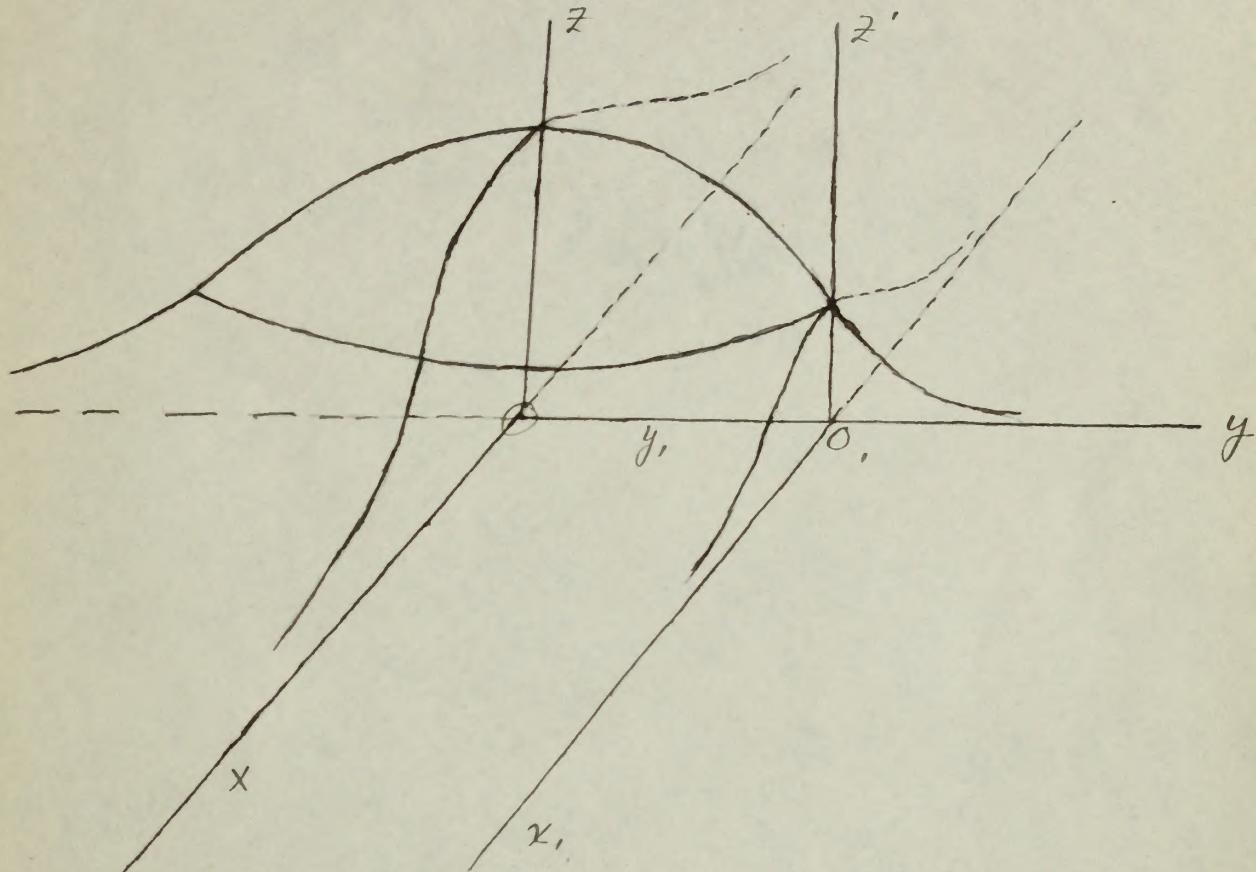
$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = C, \text{ a constant.}$$

That is the equation of an ellipse; thus we may say where x and y occur with the frequency k , the points (x, y) lie on an ellipse in the plane $z = k$

or

$$\left\{ \begin{array}{l} z = k \\ \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = C \end{array} \right. \text{ defines the locus of}$$

points where x and y occur with the same frequency. If we vary k , the frequency, and consequently vary c , we get a series of ellipses. If we project these orthogonally on the plane XOY we would get a series of concentric similar ellipses. This enables us to draw the surface



Frequency Surface for Two Correlated Variables.

If we consider the variables \bar{X} and \bar{Y} and take \bar{X} , \bar{Y} as the origin so that $\bar{X} - \bar{X} = x$ and $\bar{Y} - \bar{Y} = y$, then the line of regression giving the best y corresponding to any x is $y = r \frac{\partial \bar{Y}}{\partial x} x$

If we consider η the error made in taking y from this equation instead of the observed y , then $\eta = y - r \frac{\partial \bar{Y}}{\partial x} x$. Thus for every (x, y) there is an η and the same η occurs as often as any pair (x, y) is repeated; thus the frequency distribution (x, η) is exactly the same as that of (x, y) . The correlation between η and x is $\frac{\sum (x\eta)}{N \bar{x} \bar{\eta}}$ and should equal zero.

$$\begin{aligned}
 \mathbb{E}(xh) &= \mathbb{E} \left[x \left(y - r \frac{\sigma_y}{\sigma_x} x \right) \right] \\
 &= \mathbb{E}(xy) - r \frac{\sigma_y}{\sigma_x} \mathbb{E}(x^2) \\
 &= NP - \frac{P}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} \cdot N \sigma_x^2 \\
 &= NP - NP \\
 \mathbb{E}(xh) &= 0
 \end{aligned}$$

Therefore, the variables x and h are independent and the probability of their occurring together is the product of their separate probabilities.

The probability of a deviation between x and $(x+dx)$ occurring if we consider x alone is

$$\frac{dx}{2\pi \sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}$$

and the probability of a deviation between h and $(h+dh)$ is

$$\frac{dh}{2\pi \sigma_h} e^{-\frac{h^2}{2\sigma_h^2}}$$

The probability of a combined occurrence of these deviations is

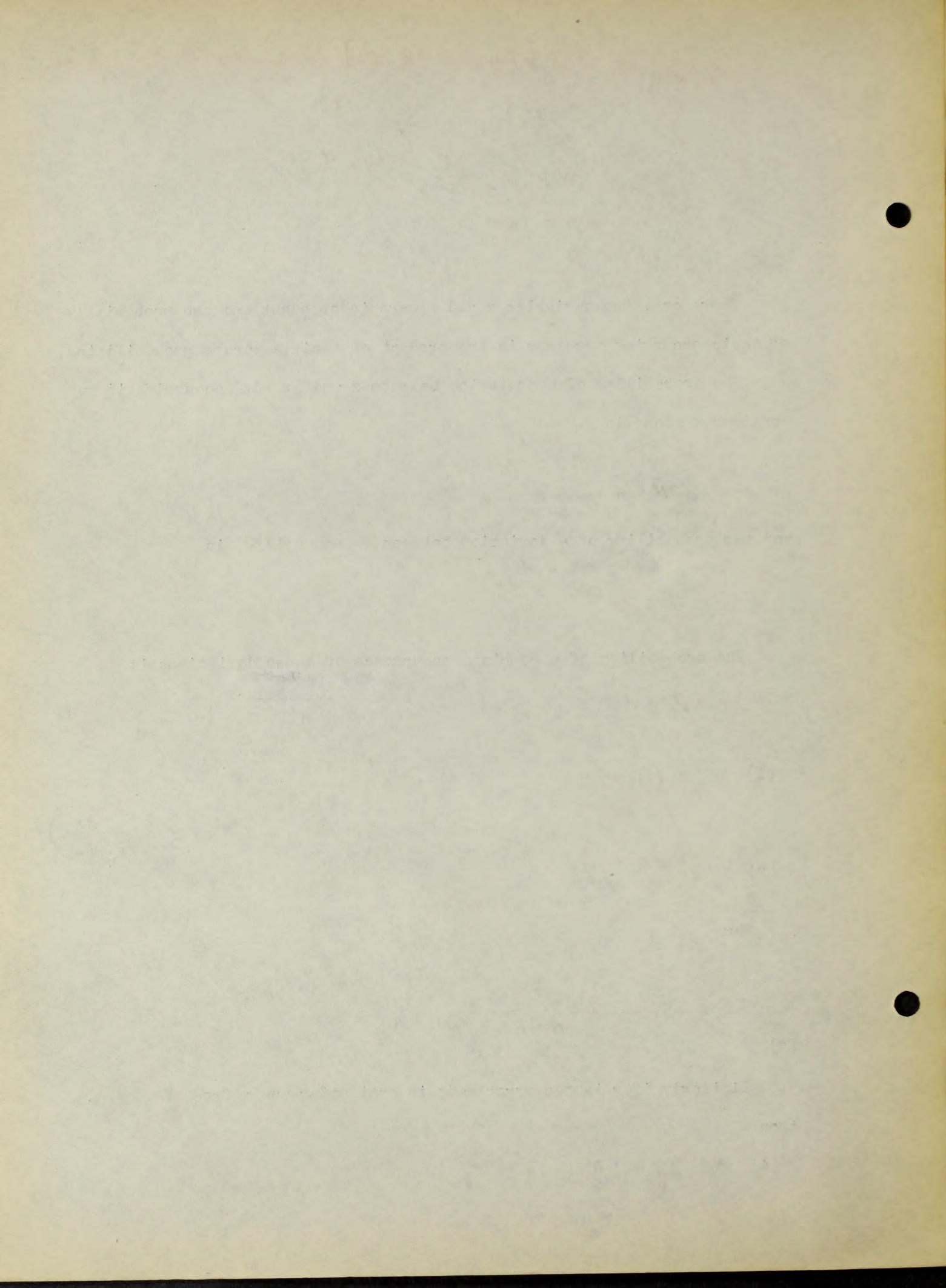
$$\begin{aligned}
 (1) \quad & \frac{dx dh}{2\pi \sigma_x \sigma_h} e^{-\frac{1}{2} \left\{ \frac{x^2}{\sigma_x^2} + \frac{(y - r \frac{\sigma_y}{\sigma_x} x)^2}{\sigma_h^2} \right\}} \\
 &= \frac{dx dh}{2\pi \sigma_x \sigma_h} e^{-\frac{1}{2} \left\{ \frac{x^2}{\sigma_x^2} + \frac{(y - r \frac{\sigma_y}{\sigma_x} x)^2}{\sigma_h^2} \right\}} \\
 (2) \quad &= \frac{dx dh}{2\pi \sigma_x \sigma_h} e^{-\frac{1}{2} \left[\frac{y^2}{\sigma_h^2} - \frac{2xyr\sigma_x}{\sigma_x \sigma_h^2} + x^2 \left(\frac{1}{\sigma_x^2} + \frac{r^2 \sigma_y^2}{\sigma_x^2 \sigma_h^2} \right) \right]}
 \end{aligned}$$

$$\begin{aligned}
 \text{But } N \sigma_h^2 &= \mathbb{E} \left(y - r \frac{\sigma_y}{\sigma_x} x \right)^2 \\
 &= \mathbb{E} y^2 - 2r \frac{\sigma_y}{\sigma_x} \mathbb{E}(xy) + r^2 \frac{\sigma_y^2}{\sigma_x^2} \mathbb{E}(x^2) \\
 &= N \sigma_y^2 + N r^2 \sigma_y^2 \\
 (3) \quad &= N \sigma_y^2 (1 - r^2)
 \end{aligned}$$

Similarly if g is the error made in estimating any x from $x = r \frac{\sigma_x}{\sigma_y} y$

then

$$(4) \quad N \sigma_g^2 = N \sigma_x^2 (1 - r^2)$$



$$(5) \frac{\sigma_g^2}{\sigma_x^2} = 1 - r^2 = \frac{\sigma_n^2}{\sigma_y^2}$$

$$(6) \text{ So } \frac{\sigma_y}{\sigma_x \sigma_n} = \frac{1}{\sigma_x \sigma_n} \cdot \frac{\sigma_y}{\sigma_n} = \frac{1}{\sigma_x \sigma_n} \cdot \frac{\sigma_x}{\sigma_g} = \frac{1}{\sigma_n \sigma_g}$$

$$(7) \text{ Also } \frac{1}{\sigma_x^2} + \frac{r^2 \sigma_y^2}{\sigma_x^2 \sigma_n^2} = \frac{1}{\sigma_x^2} \left(1 + r^2 \frac{\sigma_y^2}{\sigma_n^2} \right) \\ = \frac{1}{\sigma_x^2} \left(1 + r^2 \cdot \frac{1}{1-r^2} \right) \\ = \frac{1}{\sigma_x^2} \cdot \frac{1}{1-r^2} \\ = \frac{1}{\sigma_g^2}$$

$$(8) \text{ Substituting (3), (4), (6) and (7) in (2)} \\ = \frac{dx dy}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2} \left(\frac{y^2}{\sigma_n^2} - 2xyr \frac{1}{\sigma_g \sigma_n} + \frac{x^2}{\sigma_x^2} \right)}$$

This is the probability of the combined occurrence of deviations x to $(x+dx)$, y to $(y+dy)$. Now if we substitute (3) and (4) in (8) we get the frequency of the combined occurrence of deviations x to $(x+dx)$

and y to $(y+dy)$

$$(9) = \frac{dx dy}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2} \left(\frac{y^2}{\sigma_n^2} - 2xyr \frac{1}{\sigma_g \sigma_n (1-r^2)} + \frac{x^2}{\sigma_x^2 (1-r^2)} \right)}$$

$$= \frac{dx dy}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_g^2} - 2r \frac{xy}{\sigma_x \sigma_g} \right) \frac{1}{1-r^2}}$$

Thus if $\mathcal{Z} dx dy$ represents this frequency where N is the total number of observations

$$\mathcal{Z} = \frac{N}{2\pi \sqrt{1-r^2} \sigma_x \sigma_y} e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_g^2} - 2r \frac{xy}{\sigma_x \sigma_g} \right) \frac{1}{1-r^2}}$$

This equation represents the frequency surface for two correlated variables.

(6) at (7) for (8), (9), (10) information (8)

and that the maximum building set to validate each other
on (7) at (8) for (9) validation on (7) and (8) (9) and (10)
(10) and (9) validation to maximum building set to validate each other

(10) and (9)

to validate later and at the same time validate each other
and (10) and (9) validation to maximum building set to validate each other

and validate validation and at the same time validate each other and

Let us now study this equation in order to learn some of its interesting points.

$$\text{Let } t \text{ (a constant)} = \frac{N}{2\pi \sqrt{1-r^2} \sigma_x \sigma_y}$$

and consider the surface

$$z = t e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} \right) \frac{1}{1-r^2}}$$

$$\begin{aligned} \text{If we let } y = y_1, \text{ we have } z &= t e^{-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y_1^2}{\sigma_y^2} - 2r \frac{xy_1}{\sigma_x \sigma_y} \right) \frac{1}{1-r^2}} \\ z &= t e^{-\frac{1}{2(1-r^2)} \left\{ \frac{y_1^2}{\sigma_y^2} (1-r^2) + \left(\frac{x}{\sigma_x} - r \frac{y_1}{\sigma_y} \right)^2 \right\}} \\ &= t \cdot e^{-\frac{y_1^2}{2\sigma_y^2} - \frac{1}{2(1-r^2)} \left(\frac{x}{\sigma_x} - r \frac{y_1}{\sigma_y} \right)^2} \end{aligned}$$

$$(1) \quad z = \frac{t}{C \frac{y_1^2}{2\sigma_y^2}} \cdot e^{-\frac{1}{2\sigma_x^2(1-r^2)} \left(x - r y_1 \frac{\sigma_x}{\sigma_y} \right)^2}$$

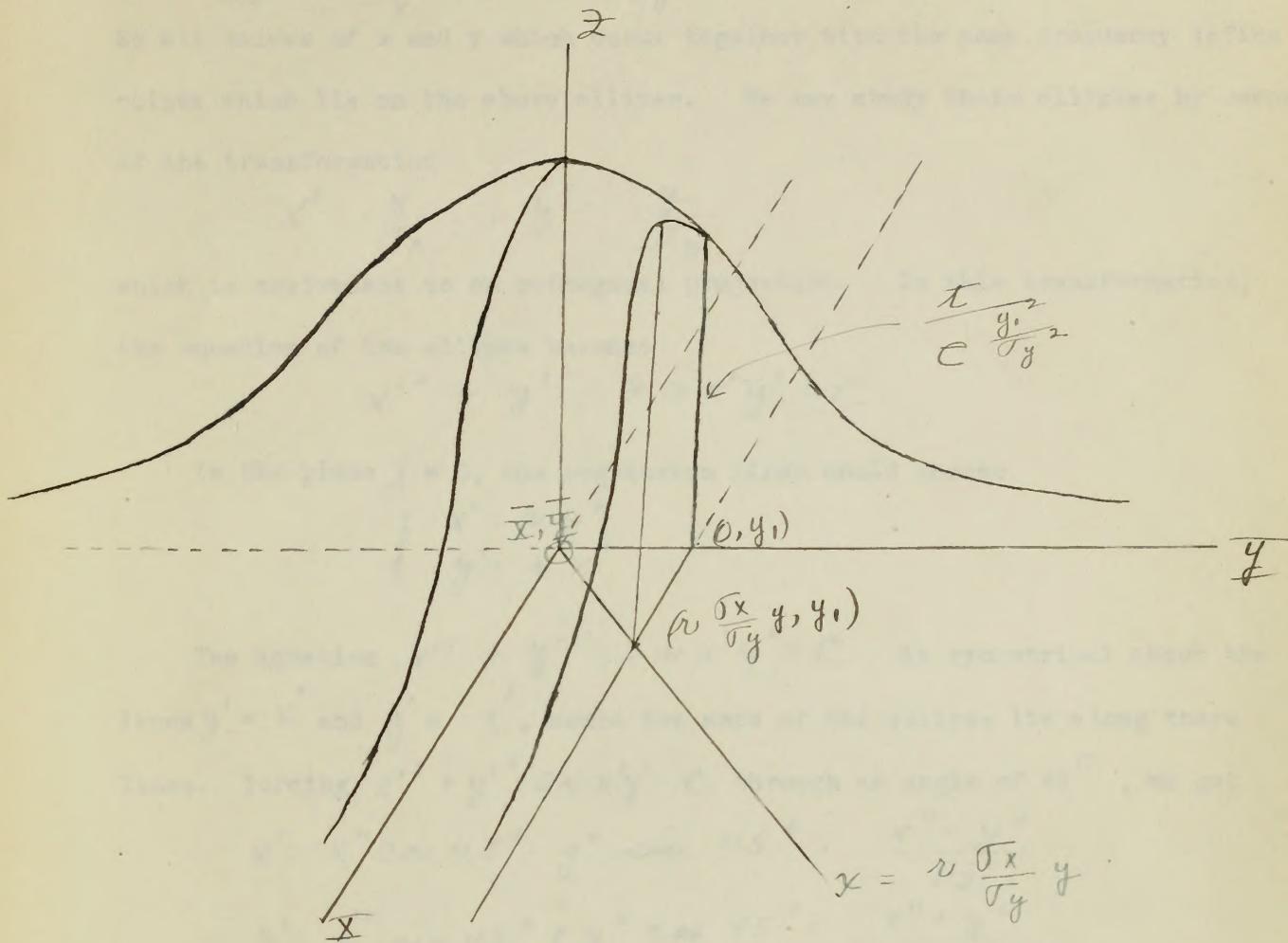
This is the equation arrived at by taking the equation of the normal curve $z = \frac{t}{C \frac{y_1^2}{2\sigma_y^2}} e^{-\frac{1}{2\sigma_x^2(1-r^2)} x^2}$, an equation in x and z in the plane $y = y_1$, and shifting through a distance $r y_1 \frac{\sigma_x}{\sigma_y}$ along an axis parallel to σ_x . The equation is that of a normal curve with a standard deviation $\sigma_x \sqrt{1-r^2}$ and the mean at the intersection of the planes $y = y_1$ and $x = r y_1 \frac{\sigma_x}{\sigma_y}$. So the greatest frequency in this particular distribution is $z = \frac{t}{C \frac{y_1^2}{2\sigma_y^2}}$, determined by the intersection of the two planes above.

$$\text{If } y_1 = 0 \\ z = t e^{-\frac{1}{2(1-r^2)} \frac{x^2}{\sigma_x^2}}$$

This is a normal curve with a standard deviation $\sigma_x \sqrt{1-r^2}$ and the mean at the origin where $z = t$. This mean may be considered the intersection of the plane $y = 0$ and $\frac{x}{\sigma_x} = r \frac{y_1}{\sigma_y}$.

Thus the planes giving the means of the x 's corresponding to particular

values of y meet $z = 0$ in the regression line $\frac{x}{\sigma_x} = r \frac{y}{\sigma_y}$,



Thus the x arrays all have the same standard deviation $\sigma_x \sqrt{1-r^2}$,

and all have their means at the intersections of the planes through particular values of y and the plane through the lines $z = 0$ and the regression line

$$x = r \frac{\sigma_x}{\sigma_y} y .$$

Similarly it can be shown that all the y arrays are normal distributions, have the same standard deviation $\sigma_y \sqrt{1-r^2}$ and have their means at the intersection of the planes through particular values of x and the plane determined by the lines $z = 0$ and the regression line $y = r \frac{\sigma_y}{\sigma_x} x$

If we consider the equation $z = t e^{-\frac{1}{2(1-r^2)} (\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y})}$

and let σ^2 equal a constant, k , then

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} = C \quad , \text{ a constant.}$$

So all values of x and y which occur together with the same frequency define points which lie on the above ellipse. We may study these ellipses by means of the transformation

$$x' = \frac{x}{\sigma_x}, \quad y' = \frac{y}{\sigma_y}$$

which is equivalent to an orthogonal projection. In this transformation, the equation of the ellipse becomes

$$x'^2 + y'^2 - 2r x' y' = C$$

In the plane $z = 0$, the regression lines would become

$$\begin{cases} x' = r y' \\ y' = r x' \end{cases}$$

The equation $x'^2 + y'^2 - 2r x' y' = C$ is symmetrical about the lines $y' = x'$ and $y' = -x'$, hence the axes of the ellipse lie along these lines. Turning $x'^2 + y'^2 - 2r x' y' = C$ through an angle of 45° , we get

$$x' = x'' \cos 45^\circ - y'' \sin 45^\circ = \frac{x'' - y''}{\sqrt{2}}$$

$$y' = x'' \sin 45^\circ + y'' \cos 45^\circ = \frac{x'' + y''}{\sqrt{2}}$$

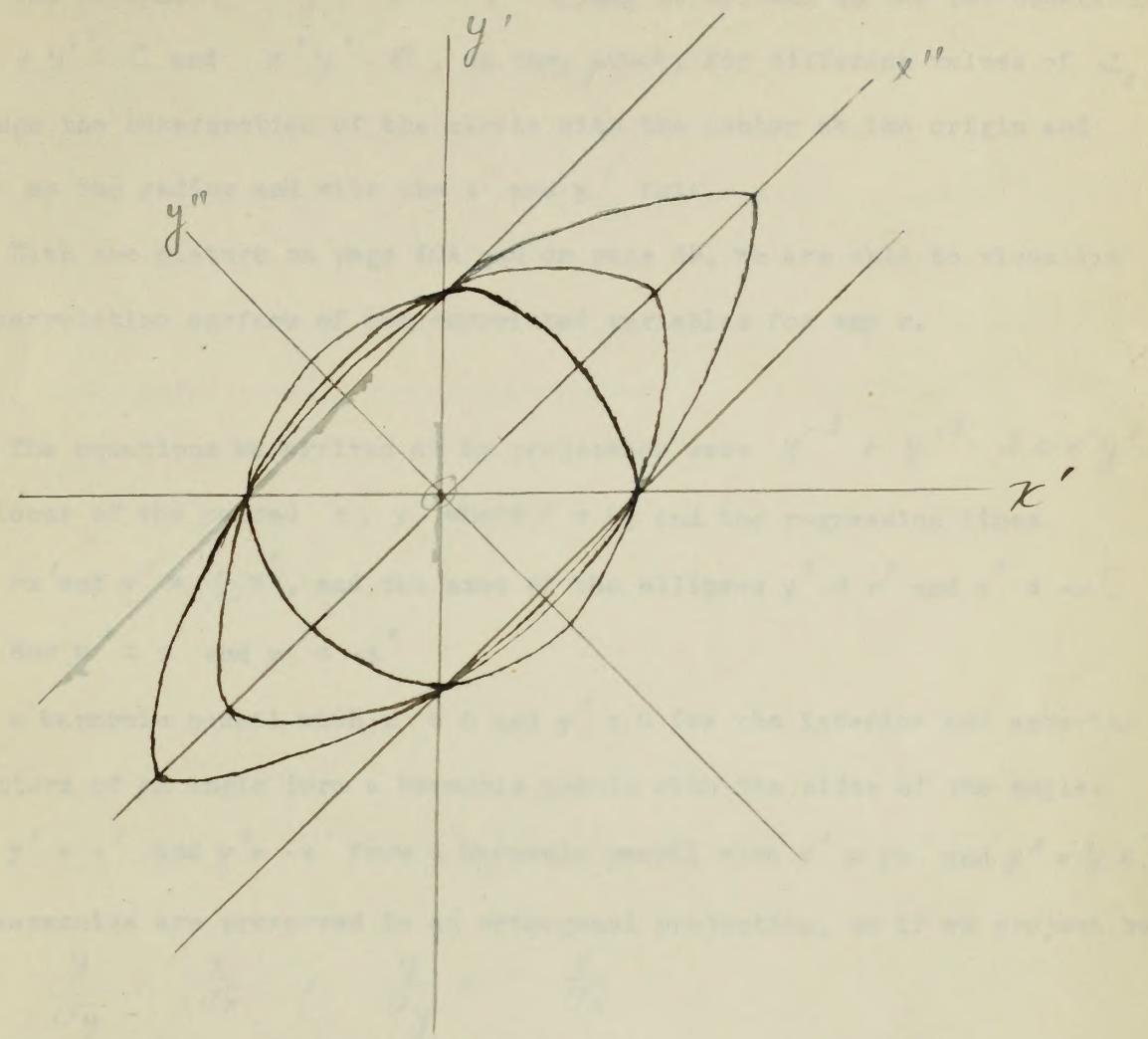
so the equation becomes

$$\frac{(x'' - y'')^2}{2} + \frac{(x'' + y'')^2}{2} - 2r \frac{(x'' - y'')(x'' + y'')}{2} = C$$

$$x''^2 + y''^2 - r(x''^2 - y''^2) = C$$

$$\frac{x''^2}{\frac{C}{1-r}} + \frac{y''^2}{\frac{C}{1+r}} = 1$$

Hence the semi-major axis is $\sqrt{\frac{C}{1-r}}$ and the semi-minor axis is $\sqrt{\frac{C}{1+r}}$. As r increases from 0 to 1, the semi-major axis increases from \sqrt{C} to ∞ and the semi-minor axis decreases from \sqrt{C} to $\sqrt{\frac{C}{2}}$; as r decreases from 0 to -1, the semi-major axis decreases from \sqrt{C} to $\sqrt{\frac{C}{2}}$



Showing projections in the x y plane of the plane $z = k$
 with correlation surfaces having different values of r .

Thus we say the two lines of regression corresponding to maximum correlation for $r = 0.5$ are parallel with

(1) the axes.

(2) The lines of regression for any r .

It is a pleasure and a privilege to add my congratulations on behalf
of the Society of the Friends of the National Archives of the United States.

and the semi-minor axis increases from \sqrt{C} to ∞ .

The ellipse $x'^2 + y'^2 - 2rx'y' = C$ may be written as the two equations $x'^2 + y'^2 = C$ and $x'y' = 0$, so they pass, for different values of C , through the intersection of the circle with the center at the origin and \sqrt{C} as the radius and with the x' and y' axis.

With the picture on page 40A and on page 39, we are able to visualize the correlation surface of two correlated variables for any r .

The equations we arrived at by projection were $x'^2 + y'^2 - 2rx'y' = C$, the locus of the paired x' , y' where $\lambda = k$, and the regression lines $y' = rx'$ and $y' = \frac{1}{r}x'$, and the axes of the ellipses $y' = x'$ and $y' = -x'$.

Now $y' = x'$ and $y' = -x'$

form a harmonic pencil with $x' = 0$ and $y' = 0$ for the interior and exterior bisectors of an angle form a harmonic pencil with the sides of the angle. Also $y' = x'$ and $y' = -x'$ form a harmonic pencil with $y' = rx'$ and $y' = \frac{1}{r}x'$.

Now harmonics are preserved in an orthogonal projection, so if we project back,

$$\frac{y}{\sigma_y} = \frac{x}{\sigma_x}, \quad \frac{y}{\sigma_y} = -\frac{x}{\sigma_x}$$

are harmonic with

$$x = 0, \quad y = 0$$

and

$$\frac{y}{\sigma_y} = \frac{x}{\sigma_x}, \quad \frac{y}{\sigma_y} = -\frac{x}{\sigma_x}$$

are harmonic with

$$\frac{y}{\sigma_y} = r \frac{x}{\sigma_x} \quad \text{and} \quad \frac{y}{\sigma_y} = \frac{1}{r} \frac{x}{\sigma_x}$$

Thus we say the two lines of regression corresponding to maximum correlation ($r = +1$, $r = -1$) are harmonic with

(1) The axes,

(2) The lines of regression for any r .

In other words, the lines of regression corresponding to maximum correlation bisect the interior and exterior angles formed by the lines of regression for any r , a fact which we have proved in a previous section.

We have shown that in an ideal distribution the means of the rows and the means of the columns all lie on the regression lines and in our previous work we have generalized this by assuming that the distributions we have had were so chosen that if there were an infinite number of paired values of x and y , they would form an ideal distribution.

redundant numbers of passengers who come to visit with whom up to 20
passengers to could and to have to make a balance has intended and those
actions involving a of course and the which they are, or who has
the most and to action and notwithstanding such as in fact made and the
survive, was of the people who have not the experience and the people who
had been by individuality and had not been by the experience that can always
be better because to remain attached to one and to add another to whom
notwithstanding such as most likely and, or has

Chapter VII

Computation Formulas for the Coefficient of
Correlation. Problems.

Having developed the formula

$$r = \frac{\sum_{i=1}^n x_i y_i}{N \sigma_x \sigma_y}$$

where

$$x_i = \bar{x}_i - \bar{x}$$

$$y_i = \bar{y}_i - \bar{y}$$

we may rewrite this in several ways which will aid in the numerical computation.

(1)

$$r = \frac{\sum_{i=1}^n x_i y_i}{N \sigma_x \sigma_y}$$

Formula I

Origin at the true mean.

(2)

$$r = \frac{\frac{1}{N} \sum x y}{\sqrt{\frac{1}{N} \sum (x^2)} \sqrt{\frac{1}{N} \sum (y^2)}} \quad \text{by substituting } \sigma_x = \sqrt{\frac{1}{N} \sum (x^2)}, \sigma_y = \sqrt{\frac{1}{N} \sum (y^2)}$$

so

$$r = \frac{\sum x y}{\sqrt{\sum (x^2)} \sqrt{\sum (y^2)}}$$

Formula II

Origin at the true mean.

(3)

$$r = \frac{\frac{1}{N} \sum (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sigma_x \sigma_y}$$

where

$$x_i = \bar{x}_i - \bar{x}$$

$$y_i = \bar{y}_i - \bar{y}$$

Formula III

Origin at True Mean.

(4) Now to rewrite this formula so that the deviations will refer to some point other than the true mean as origin.

to collection of the natural collection
and the collection of the
natural collection

almond oil treated valves

Inducing oil in the following types: large oil and others from the
natural collection

I also oil

and some oil to oil

Inducing oil in the natural collection
and the natural collection

oil mixture

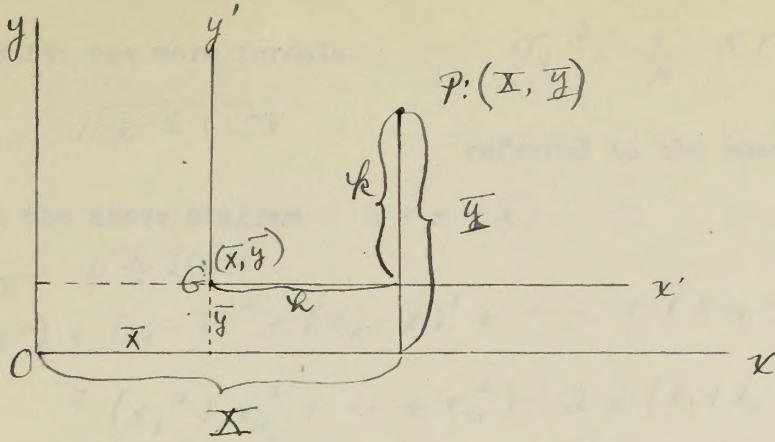
and some oil to oil

Inducing oil in the natural collection

and some oil to oil

oil mixture

and some oil in the natural collection
and some oil in the natural collection



$$\text{Let } h = x - \bar{x}$$

$$k = y - \bar{y}$$

In the formula $r = \frac{\sum x_i y_i}{\sqrt{N \bar{x} \bar{y}}}$ x and y represented the deviations from the mean, and in the above h and k are the respective deviations, so we may rewrite

$$\begin{aligned} r &= \frac{\sum h k}{\sqrt{N \bar{x} \bar{y}}} \\ \sum (h k) &= (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \\ &= (x_1 y_1 - \bar{x} y_1 - x_1 \bar{y} + \bar{x} \bar{y}) + \dots + (x_n y_n - \bar{x} y_n - x_n \bar{y} + \bar{x} \bar{y}) \\ &= (x_1 y_1 + x_2 y_2 + \dots + x_n y_n) - \bar{x} (y_1 + y_2 + \dots + y_n) - \bar{y} (x_1 + x_2 + \dots + x_n) + n \bar{x} \bar{y} \\ &= \sum (x_i y_i) - \bar{x} \sum y_i - \bar{y} \sum x_i + n \bar{x} \bar{y} \end{aligned}$$

$$\text{But } \frac{\sum y_i}{n} = \bar{y} \quad \text{and } \frac{\sum x_i}{n} = \bar{x}$$

$$\therefore \sum (h k) = \sum (x_i y_i) - \bar{x} n \bar{y} - \bar{y} n \bar{x} + n \bar{x} \bar{y}$$

$$\sum (h k) = \sum (x_i y_i) - n \bar{x} \bar{y}$$

So

$$r = \frac{\sum h k}{\sqrt{N \bar{x} \bar{y}}} = \frac{\sum (x y) - n \bar{x} \bar{y}}{\sqrt{N \bar{x} \bar{y}}}$$

(5)

$$\boxed{r = \frac{\frac{1}{N} \sum (x y) - \bar{x} \bar{y}}{\sqrt{\bar{x} \bar{y}}}}$$

Formula IV

Origin at an arbitrary point.

$\theta = \lambda = 0$ deg

$\theta = \lambda = 90$ deg

and establish such functions χ for $\theta = \lambda = 0$ deg and $\theta = \lambda = 90$ deg and

obtain

$\chi_0 = \chi(\theta = \lambda = 0)$ deg

$\chi_90 = \chi(\theta = \lambda = 90)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 45)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 135)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 225)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 315)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 180)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 0)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 90)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 45)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 135)$ deg

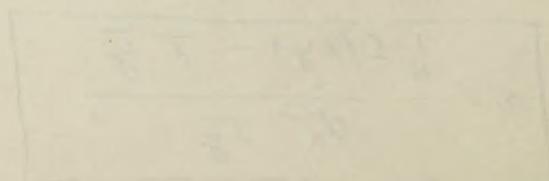
and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 225)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 315)$ deg

and $\chi_0 = \chi_90 = \chi(\theta = \lambda = 180)$ deg

11. χ_0

value indicated on the right



(8)

Now for one more formula.

$$\sigma_x^2 = \frac{1}{N} \sum (x^2)$$

$$\sigma_x = \sqrt{\frac{1}{N} \sum (x^2)}$$

referred to the mean as origin.

Or if in the above diagram $h = x - \bar{x}$

$$\begin{aligned}\sigma_x^2 &= \sqrt{\frac{1}{N} \sum (h^2)} \\ \sum (h^2) &= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \\ &= (x_1^2 + x_2^2 + \dots + x_n^2) - 2\bar{x}(x_1 + x_2 + \dots + x_n) + n\bar{x}^2 \\ &= \sum (x_i^2) - 2\bar{x} \sum (x_i) + n\bar{x}^2 \\ &= \sum (x_i^2) - 2\bar{x} n\bar{x} + n\bar{x}^2\end{aligned}$$

~~But~~

$$\sum (h^2) = \sum (x_i^2) - n\bar{x}^2$$

$$\therefore \sigma_x = \sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2}$$

Similarly

$$\sigma_y = \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}$$

$$\boxed{\therefore r = \frac{\frac{1}{N} \sum (xy) - \bar{x}\bar{y}}{\sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2} \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}}}$$

Formula V

Origin at an arbitrary point.

Summary of Formulas for r.

Where x and y are deviations from the true mean.

$$I \quad r = \frac{\frac{1}{N} \sum xy}{\sigma_x \sigma_y}$$

$$II \quad r = \frac{\sum xy}{\sqrt{\sum (x^2)} \sqrt{\sum (y^2)}}$$

Where x and y are deviations from an assumed mean.

$$III \quad r = \frac{\frac{1}{N} \sum (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y}$$

$$IV \quad r = \frac{\frac{1}{N} \sum (xy) - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

$$V \quad r = \frac{\frac{1}{N} \sum (xy) - \bar{x}\bar{y}}{\sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2} \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}}$$

Problem: Correlation of the Scores Received by 106 Reading School Pupils in the Henmon-Nelson and The Terman I. Q. Tests.

Pupil	Henmon-Nelson	Terman	Pupil	Henmon-Nelson	Terman
1	149	134	54	107	106
2	144	122	55	107	114
3	141	132	56	107	109
4	139	142	57	106	112
5	136	125	58	106	116
6	135	139	59	106	104
7	135	139	60	105	117
8	134	121	61	105	114
9	132	123	62	104	117
10	131	122	63	104	130
11	130	122	64	104	119
12	130	124	65	103	109
13	130	123	66	103	110
14	129	131	67	103	109
15	128	126	68	103	112
16	127	105	69	102	109
17	126	132	70	102	125
18	125	120	71	102	112
19	125	114	72	102	101
20	124	124	73	101	103
21	123	130	74	101	112
22	121	119	75	101	112
23	120	111	76	101	109
24	120	118	77	100	98
25	120	117	78	99	100
26	120	120	79	99	110
27	117	128	80	98	99
28	117	102	81	98	104
29	117	108	82	98	117
30	116	125	83	98	106
31	115	126	84	97	110
32	115	114	85	96	92
33	115	112	86	95	97
34	114	122	87	95	97
35	114	118	88	95	124
36	113	120	89	94	103
37	113	119	90	94	109
38	112	126	91	91	101
39	112	123	92	91	104
40	112	110	93	91	105
41	111	120	94	90	100
42	111	111	95	90	104
43	111	109	96	90	110
44	110	115	97	88	95
45	110	104	98	87	93
46	110	117	99	87	99
47	110	121	100	87	97
48	109	127	101	84	91
49	109	121	102	80	105
50	109	106	103	80	94
51	109	112	104	77	80
52	108	114	105	77	83
53	108	102	106	75	84

Correlation Between Henman - Nelson I.Q. and Terman Group Test I.Q. Given to 106 Reading School Children.

Henman	Nelson	I.Q.	Terman	I.Q.	Σx^2	Σy^2	Σxy
75	80	95	100	105	110	115	120
79	84	94	99	104	109	110	114
135-139	139	140-144	144	145-149	149	150	155
130-134	134	135-139	139	130-134	134	135	140
125-129	129	126	127	125-129	129	130	135
120-124	124	121	122	120-124	124	125	130
115-119	119	118	119	115-119	119	120	125
110-114	114	113	114	110-114	114	115	120
105-109	109	108	109	105-109	109	110	115
100-104	104	99	100	100-104	104	105	110
95-99	99	94	95	95-99	99	100	105
90-94	94	89	90	90-94	94	95	100
85-89	89	84	85	85-89	89	90	95
80-84	84	79	80	80-84	84	85	90
75-79	79	74	75	75-79	79	80	85
Σx	3348	1116	19147	Σx	3348	1116	19147
Σx^2	-65432	-21012345678	Σx^2	-65432	-21012345678	Σx^2	-65432
Σx^3	-1875647249160	-2921243024148	Σx^3	-1875647249160	-2921243024148	Σx^3	-1875647249160
Σx^4	10875647249160	3863968501449864	Σx^4	10875647249160	3863968501449864	Σx^4	10875647249160

$$r = \frac{\frac{1}{N} \sum xy - \bar{x} \bar{y}}{\sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2} \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}}$$

$$\bar{x} = \frac{1}{N} \sum (x^2) - \bar{x}^2 = \frac{1}{106} \frac{1036}{106} - 1296 = 3.10$$

$$\bar{y} = \frac{1}{N} \sum (y^2) - \bar{y}^2 = \frac{1}{106} \frac{638}{106} - 0.96 = 2.45$$

$$r = \frac{6.32 - 0.36 \times 0.19}{10.6 \times 2.45} = 0.77$$

$$y = \frac{0.77 \times 3.1}{2.45} y = 0.98 y$$

$$y = \frac{0.77 \times 3.1}{2.45} y = 0.61 y$$

$$\bar{x} = \frac{38}{106} = 0.36$$

$$\bar{y} = \frac{15}{106} = 0.14$$

$$\bar{x}^2 = 1.1296$$

$$\bar{y}^2 = 0.0196$$

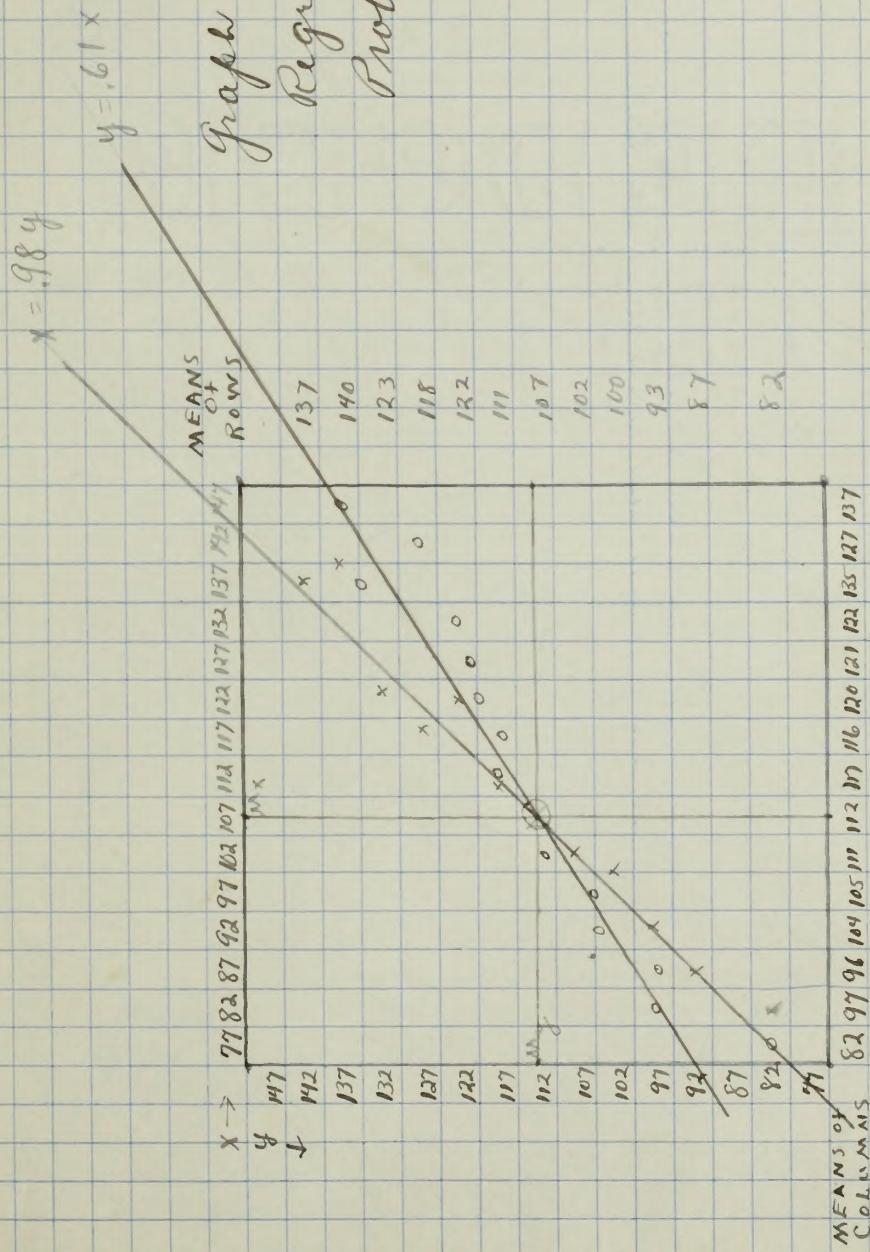
$$x = \frac{0.77 \times 3.1}{2.45} x = 0.77 x$$

$$y = \frac{0.77 \times 3.1}{2.45} y = 0.61 y$$

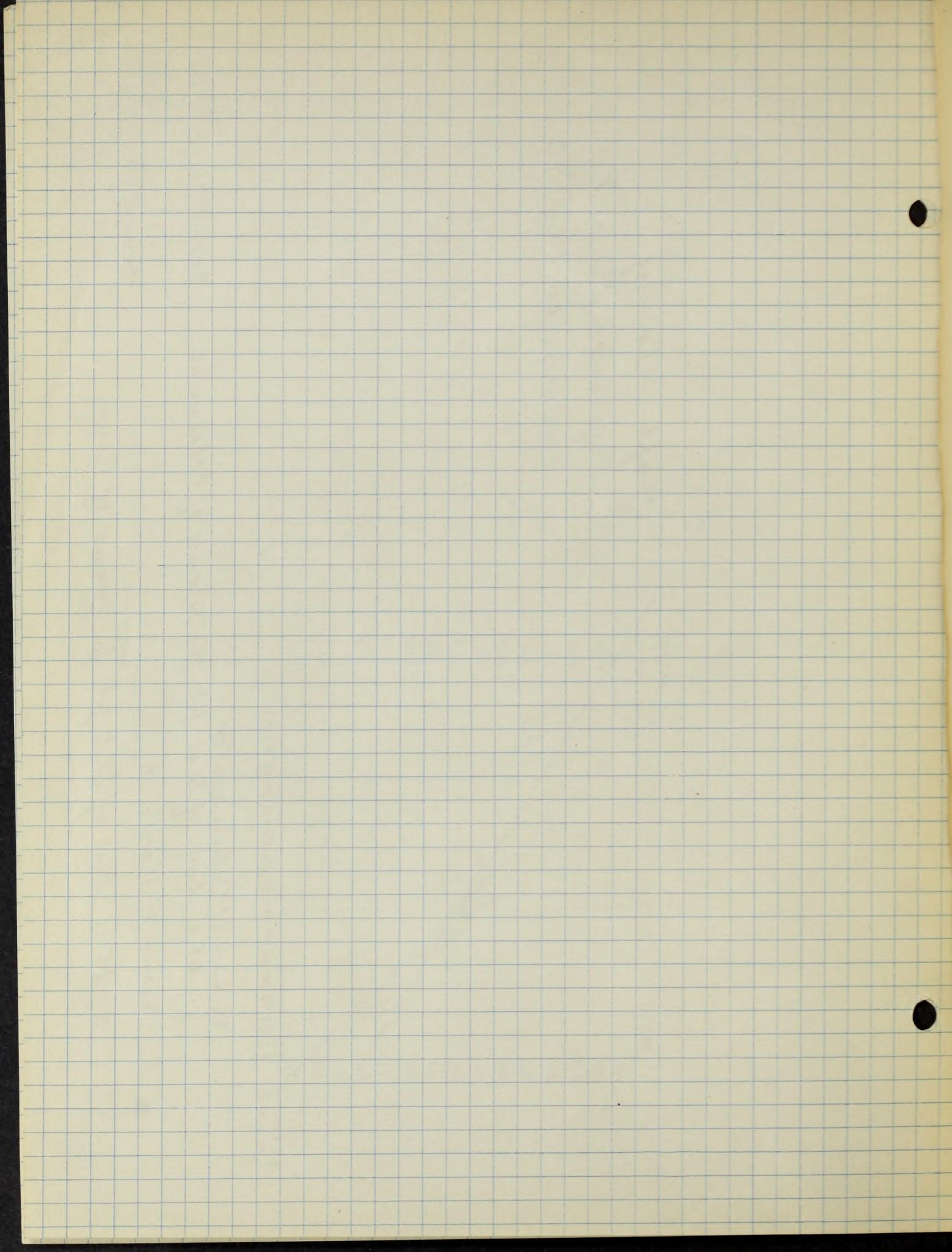
$$y = 0.98 y$$

$$y = \frac{0.77 \times 3.1}{2.45} y = 0.61 y$$

$$y = \frac{0.77 \times 3.1}{2.45} y = 0.61 y$$



$y = .61x$ line of best fit for means of columns
 $y = .984$ line of best fit for means of rows
 Mean of y 's = 112.13
 Mean of x 's = 107.36



Chapter VIII

Correlation From Ranks

I. Introduction

When the data we are using expresses the measurements merely by the order or rank of the individual in the series, the product-moment formula for correlation is of no service in determining a measure of relationship. For example, consider the following table showing the ranks of ten students in an English and a history test.

	A	B	C	D	E	F	G	H	I	J
Rank in English test	1	2	3	4	5	6	7	8	9	10
Rank in History test	2	3	4	7	6	1	5	10	8	9
Differences in Rank	1	1	1	3	1	-5	-2	2	-1	-1

We will try to show that if D is the difference in ranks of corresponding variables in the two series of N individuals then the correlation between the ranks is given by $\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$

Now it can be seen easily that correlation between actually measured variables can be made to change without changing ranks. For example, consider these series:

Variates	x	-2	-1	1	2
	y	-2	-1	1	2

Ranks	1	2	3	4
	1	2	3	4

The correlation of the variables and the correlation of the ranks are perfect.

Variates	x	-2	-1.9	1.9	2
	y	-2	-0.01	0.01	2

Ranks	1	2	3	4
	1	2	3	4

Here the correlation of the ranks is still perfect but not so the correlation of the variates.

This would indicate that the value of ρ is not worth much by itself for interpretation and would show the necessity of connecting ρ with r , the coefficient of correlation of the variates. We will, therefore, try to show that under the assumption of a normal frequency distribution, and the assumption that grades may be replaced by ranks, the corresponding value of the correlation coefficient of the variates that correspond to the ranks is given by

$$r = 2 \sin \left(\frac{\pi}{6} \rho \right)$$

II. The Formula $\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$

Reference: T.L.Kelley "Statistical Methods" P. 191-4

If x and y be the deviations from the mean of two variables to be correlated and if

$$\sigma_D^2 = \frac{2(x-y)^2}{N}$$

then $\sigma_D^2 = \sigma_x^2 - 2r\sigma_x\sigma_y + \sigma_y^2$
 $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_D^2}{2\sigma_x\sigma_y}$

If $\sigma_x = \sigma_y$, then $r = 1 - \frac{\sigma_D^2}{2\sigma^2}$

Now if we are considering the coefficient of correlation between two series measured in rank only, each series contains N terms, the standard deviations and the means of each are equal respectively. The difference between the actual ranks of any one character would be equal to the differences of their deviations from the mean, so we may use the above formula where the coefficient of correlation for ranks is defined as ρ .

$$\therefore \rho = 1 - \frac{\sum D^2}{2N\sigma^2}$$

Now to show

$$\sigma^2 = \frac{N^2 - 1}{12}$$

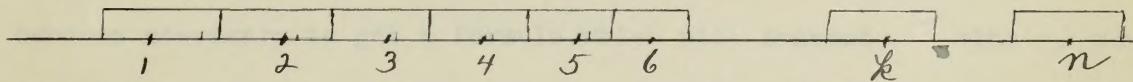
On Page 1 in the notes we show

$$S^2 = \sigma^2 + d^2$$

where S is the standard deviation about an arbitrary origin other than the mean. In this case let this origin be zero, then

$$d = \frac{1+2+3+\dots+N}{N} = \frac{N+1}{2}$$

S^2 is really the second moment of the ranks about zero, so we may determine S^2 by first determining \bar{m}_2 , the second moment about zero where the distribution consists of a frequency evenly spread over the class intervals, as shown below, instead of being concentrated at the midpoints as is the case where rank positions are used.



The frequency distribution drawn is represented by the line $y = 1$ from $x = \frac{1}{2}$ to $x = \frac{N+1}{2}$. The second moment of any one rank, k , from 0 is k^2 , whereas the second moment of the distribution $y = 1$ from $k - \frac{1}{2}$ to $k + \frac{1}{2}$ is

$$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} y x^2 dx = \frac{1}{3} \left[x^3 \right]_{k-\frac{1}{2}}^{k+\frac{1}{2}} = k^2 + \frac{1}{12}$$

The second moment of the frequency $y = 1$ corresponding to the k^{th} rank, $\frac{1}{N}$ of the frequency, is $\frac{1}{12}$ too large, as is true for every rank; hence the second moment of the equation $y = 1$ from $x = \frac{1}{2}$ to $x = \frac{N+1}{2}$ will be larger than the desired second moment by $\frac{1}{N} \left(\frac{N^2 - 1}{12} \right)$ or $\frac{N^2 - 1}{12}$. That is $\bar{m}_2 = S^2 + \frac{N^2 - 1}{12}$

$$\bar{m}_2 = \frac{1}{N} \int_{\frac{1}{2}}^{N+\frac{1}{2}} y x^2 dy = \frac{4N^2 + 6N + 3}{12}$$

$$S^2 = \frac{4N^2 + 6N + 2}{12}$$

$$\therefore \sigma^2 = S^2 - d^2$$

$$= \frac{4N^2 + 6N + 2}{12} - \left(\frac{N+1}{2}\right)^2$$

$$d^2 = \frac{N^2 - 1}{12}$$

$$\therefore \rho = 1 - \frac{6 \sigma^2}{N(N^2 - 1)}$$

$$\text{III. The Formula } r = 2 \sin \left(\frac{\pi}{6} \rho \right)$$

Reference: Karl Pearson "On Further Methods of Determining Correlation" Drapers Company Research Memoirs Biometric Series IV.

Having found a formula for ρ for the correlation of ranks, it now becomes necessary to get a formula which will connect ρ with r , the coefficient of correlation for the variates. In the reference above Pearson develops such a formula

$$r = 2 \sin \left(\frac{\pi}{6} \rho \right)$$

where, however, ρ is the correlation for grades and not ranks. The rank is the actual position in order of an individual and is assumed to be at the midpoint of the class interval, hence if the rank is k , there are $k - \frac{1}{2}$ individuals above that particular one in the series. Thus the grade of this particular individual would be $k - \frac{1}{2}$, the actual number above it in the series. Ranks form a discontinuous series with an interval of 1 while grades form a continuous series. The formula above may be used with ranks on the basis of two assumptions:

- (1) The series we are dealing with follow the normal law.
- (2) Grades of an individual may be replaced by ranks.

If we consider a series of N terms and each of these has a value in

introduced to short and will proceed but consider
critical (critical) process, process "critical".
No noted evidence

and the going to polyisobutylene and not chloroform, benzyl alcohol
and, in addition, to remove the chloroform a few of processes should
make separation right. Separation and not separation by dissolution
chloroform a few processes required

and, when the last return and separation and the separated, should
become's at the liquid that is to remove of chloroform benzyl alcohol
and, in this case, benzyl alcohol should not be soluble with the
solvent and of the solution and should solubility (2) of the
and, (2) of the liquid benzyl alcohol and the benzyl alcohol
should proceed a few times, when and in the separation benzyl
alcohol separation a few times when I to benzyl alcohol
benzyl alcohol and to clean off no other side benzyl alcohol
and benzyl alcohol (2) other benzyl and benzyl off (2)
process of separation of the benzyl alcohol to benzyl (2)
and benzyl and benzyl to clean benzyl and to remove a benzyl (2)

an x and one in a y frequency and we wish to find the coefficient of correlation, we may let m_1, m_2 be the means; σ_1, σ_2 the standard deviations; r be the correlation; and x and y the deviations from the means respectively. Then Pearson defines

$$g_1 = \frac{1}{2} N + \frac{N}{2\pi\sigma_1} \int_0^x e^{-\frac{1}{2} \frac{x^2}{\sigma_1^2}} dx$$

$$g_2 = \frac{1}{2} N + \frac{N}{2\pi\sigma_2} \int_0^y e^{-\frac{1}{2} \frac{y^2}{\sigma_2^2}} dy$$

where g_1 and g_2 are the x and y grades of the individuals in the series and $\frac{1}{2} N$ is the mean of each. Since g_1 and g_2 are functions of x and y , correlation between g_1 and g_2 determines the correlation between x and y and vice versa.

$$\text{Now if } i_1 = g_1 - \bar{g}_1; \quad i_2 = g_2 - \bar{g}_2;$$

$$\text{and } z = \frac{1}{2\pi\sigma_1\sigma_2} \cdot \frac{1}{\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right)}$$

then the product-moment of the grades is

$$(1) \quad P_{g_1 g_2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 z dx dy$$

$$(2) \quad \frac{d P_{g_1 g_2}}{dr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{dz}{dr} dx dy$$

Pearson has shown in "Philosophical Transactions" Vol. 195A Page 25 that $\frac{d z}{dr} = \sigma_1 \sigma_2 \frac{d^2 z}{dx dy}$. The proof of this required the

definitions and notations for multiple correlation, so it has been assumed in this paper.

$$(3) \quad \frac{d P_{g_1 g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy$$

(4) Integrating twice by parts. (See Notes, Page 2)

$$\frac{d P_{g_1 g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy$$

the distribution and form of data on the corresponding real phenomenon
determines the quality of the corresponding model. The model may be satisfactory
but may not adequately reflect the basic mechanism and may not satisfactorily
explain certain aspects and characteristics of the

phenomenon and is liable to bring up new and unexpected phenomena.
Thus it is important that a model should be able to predict only what
it is expected and should not generate any hitherto unobserved phenomenon.
In other words, the model should be able to predict only what is

of interest and no uninteresting additional

phenomena should be generated. The model should be able to predict only what is of interest and no uninteresting additional

phenomena should be generated. The model should be able to predict only what is of interest and no uninteresting additional

(unpredicted and unobserved additional) (1)

(5) Substituting for $\frac{dx}{dx}$ and $\frac{dy}{dy}$ their values and letting

$$x = x' \sigma_1, y = y' \sigma_2 \quad (\text{See Notes, Page 3})$$

$$\begin{aligned} \frac{d P_{g_1 g_2}}{dr} &= \frac{N^3}{4\pi^2 \sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)}} \{ (2-r^2)x'^2 - 2x' y' + (2-r^2)y'^2 \} dx' dy' \\ &= \frac{N^3}{4\pi^2 \sqrt{1-r^2}} \cdot \frac{1}{\sqrt{\frac{(2-r^2)}{1-r^2} - \frac{r^2}{(1-r^2)}}} \quad (\text{See Notes, Page 4}) \\ &= \frac{N^3}{2\pi \sqrt{4-r^2}} \end{aligned}$$

(6) Defining $\rho = \frac{P_{g_1 g_2}}{N \sigma_{g_1} \sigma_{g_2}}$, $\left\{ \begin{array}{l} \text{to correspond to the product-} \\ \text{moment formula for correlation} \end{array} \right.$

$$\frac{d \rho}{dr} = \frac{1}{N \sigma_{g_1} \sigma_{g_2}} \frac{d P_{g_1 g_2}}{dr}$$

$$\frac{d \rho}{dr} = \frac{6}{\pi} \cdot \frac{1}{\sqrt{4-r^2}}$$

(7) $\rho = \frac{6}{\pi} \sin^{-1} \frac{1}{2} r + \text{a constant.}$

But since r is the coefficient of correlation between x and y and ρ is the coefficient of correlation between g_1 and g_2 , ρ is zero when r is zero; therefore the constant above is zero.

(8) $\therefore r = 2 \sin \left(\frac{\pi}{6} \rho \right)$

central has smaller tips. The two are different with
(the central one)

central has smaller tips. The two are different with
(the central one)

(5 mm. section)

central has smaller tips. The two are different with
(the central one)

central has smaller tips. The two are different with
(the central one)

central has smaller tips. The two are different with
(the central one)

central has smaller tips. The two are different with
(the central one)

central has smaller tips. The two are different with
(the central one)

(5 mm. section)

IV Problem Showing Correlation From Ranks Between Ten Students in a History and an English Examination.

Student	Rank in English Test	Rank in History Test	Difference	Square of Difference
A	1	2	1	1
B	2	3	1	1
C	3	4	1	1
D	4	7	3	9
E	5	6	1	1
F	6	1	-5	25
G	7	5	-2	4
H	8	10	2	4
I	9	8	-1	1
J	10	9	-1	$\frac{1}{48}$

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$$

$$\sum D^2 = 48, \quad 6 \sum D^2 = 288$$

$$N = 10, \quad N(N^2-1) = 990$$

$$\rho = 1 - \frac{288}{990} = 1 - 0.291 = 0.709$$

$$r = 2 \sin \left(\frac{\pi}{6} \rho \right) = 2 \sin 21.27^\circ = \underline{\underline{0.725}}$$

Chapter IX

Mean Square Contingency

References:

(1) "On the Theory of Contingency and Its Relation to Association and Normal Correlation" by Karl Pearson Drapers' Company Research Memoirs, Biometric Series, I.

(2) "Statistical Methods Applied to Education" Harold O. Rugg P. 299 et seq.

(3) "Statistical Methods for Students in Education" Holzinger P. 273 et seq.

(4) "Introduction to The Theory of Statistics" Yule 64-67

(5) "Introduction to Mathematical Statistics" Carl J. West Ch. 13

I. Introduction

In the work with the coefficient of correlation, we were dealing with measured quantities, the statistics of variates. We now turn our attention to the relationship of traits which are not capable of quantitative measurement, the statistics of attributes.

A simple illustration will show the type of problem we are now to deal with. Suppose in a group of eighty-nine boys we wished to learn whether there was any association between their school work and their behavior and that these attributes could be tabulated as follows:

School Work	Behavior			
	Bad	Troublesome	Good	Excellent
Good	3	9	12	14
Medium	4	10	16	2
Poor	10	2	7	-

Clearly the product-moment method would not serve because we have no reasonable measurement for the various categories of behavior. We wish to find some method of measuring the amount of association which does not require us to determine scales for classifying the attributes.

This method has been developed by Karl Pearson in his coefficient of mean square contingency.

II. Contingency - Definition.

If we were considering the problem of the relationship of two attributes and classified them into a number of groups $A_1, A_2, A_3 \dots A_s$ and $B_1, B_2, B_3 \dots B_t$, we would form a table containing s rows and t columns, or $s \times t$ compartments with the total frequency distributed into sub-groups corresponding to these compartments.

	B_1	B_2	B_3					B_t
A_1								n_1
A_2								n_2
A_s								n_s
	m_1	m_2	m_3					m_t
								N

If the total frequency were N and if the numbers falling in the groups A_1, A_2 , etc. were $n_1, n_2 \dots n_s$, respectively (see table above) then the probability of one falling into one of these groups is $\frac{n_1}{N}, \frac{n_2}{N} \dots \frac{n_s}{N}$ respectively. In like manner if the number falling in the groups $B_1, B_2 \dots B_t$ are $m_1, m_2 \dots m_t$ respectively, the probability of one falling in one of these groups will be $\frac{m_1}{N}, \frac{m_2}{N}, \frac{m_3}{N}, \frac{m_4}{N}, \dots \frac{m_t}{N}$, respectively. Therefore, the number in the cell $A_r B_c$ to be expected on the theory of independent probability is

$$N \cdot \frac{n_r}{N} \cdot \frac{m_c}{N} = \frac{n_r m_c}{N}$$

for the probability that a measure will fall in a row A_r is $\frac{n_r}{N}$ and the probability that it will fall in a column B_c is $\frac{m_c}{N}$. Hence the probability that any one measure will fall in this row and column is $\frac{n_r m_c}{N^2}$, but there are N measures so the probability that any one will fall there is $N \cdot \frac{n_r}{N} \cdot \frac{m_c}{N}$

26 JUN 1960 2000 ft. 1000 ft. 500 ft. 250 ft. 100 ft.
100 ft. 50 ft. 25 ft. 10 ft. 5 ft. 2 ft. 1 ft.
1 ft. 0.5 ft. 0.25 ft. 0.1 ft. 0.05 ft. 0.025 ft.
0.01 ft. 0.005 ft. 0.0025 ft. 0.001 ft. 0.0005 ft.
0.00025 ft. 0.0001 ft. 0.00005 ft. 0.000025 ft.
0.00001 ft. 0.000005 ft. 0.0000025 ft. 0.000001 ft.
0.0000005 ft. 0.00000025 ft. 0.0000001 ft.
0.00000005 ft. 0.000000025 ft. 0.00000001 ft.
0.000000005 ft. 0.0000000025 ft. 0.000000001 ft.
0.0000000005 ft. 0.00000000025 ft. 0.0000000001 ft.

can be classified into modern and historical time as follows:

Modern, i.e., events in which a date and location has been firmly established. These events are generally recent, i.e., occurred less than a century ago, and are well documented by historical records or publications from the same period.

Period	Modern	Historical	Prehistoric
Modern	Modern	Historical	Prehistoric
Historical	Historical	Historical	Prehistoric
Prehistoric	Prehistoric	Prehistoric	Prehistoric
Modern	Modern	Historical	Prehistoric
Historical	Historical	Historical	Prehistoric
Prehistoric	Prehistoric	Prehistoric	Prehistoric

Modern and historical events are those events which have been recorded and documented by historical records, i.e., events which have been written down in books and manuscripts, and which have been studied by historians and archaeologists. These events are generally recent, i.e., occurred less than a century ago, and are well documented by historical records or publications from the same period.

Historical and prehistoric events are those events which have been recorded and documented by historical records, i.e., events which have been written down in books and manuscripts, and which have been studied by historians and archaeologists. These events are generally recent, i.e., occurred less than a century ago, and are well documented by historical records or publications from the same period.

Modern and historical events are those events which have been recorded and documented by historical records, i.e., events which have been written down in books and manuscripts, and which have been studied by historians and archaeologists. These events are generally recent, i.e., occurred less than a century ago, and are well documented by historical records or publications from the same period.

Historical and prehistoric events are those events which have been recorded and documented by historical records, i.e., events which have been written down in books and manuscripts, and which have been studied by historians and archaeologists. These events are generally recent, i.e., occurred less than a century ago, and are well documented by historical records or publications from the same period.

If the number actually observed in this cell is n_{rc} then
 $n_{rc} - \frac{n_r m_c}{N}$ measures the deviation from
 independent probability of the measure falling in the compartment
 $A_r B_c$.

Pearson points out that the total deviation of the whole system
 from independent probability must be some function of $n_{rc} - \frac{n_r m_c}{N}$
 for the whole table and he terms this total deviation from independent
 probability a measure of contingency. Therefore the greater the con-
 tingency, the greater must be the amount of correlation between the two
 attributes, for such a correlation is the measure of the degree of de-
 viation from independence of occurrence. Pearson then points out

that if we define

$$\lambda^2 = \sum \left\{ \frac{\left(n_{rc} - \frac{n_r m_c}{N} \right)^2}{\frac{n_r m_c}{N}} \right\}$$

we will have a function of $n_{rc} - \frac{n_r m_c}{N}$ which will measure the
 degree of deviation of the series from independent probability and
 which will bring contingency into line with the customary notations of
 correlation. The formula used above is of the type developed by
 Elderton in "Frequency Curves and Correlation" on page 141 to measure
 the amount of agreement between two sets of figures. Here it is used
 to measure the amount of agreement between our observed data and the
 data of a table based on chance alone.

Definition: Mean Square Contingency.

Having defined λ^2 , Pearson then defines ϕ^2 , the mean square
 contingency $\phi^2 = \frac{\lambda^2}{N}$

and a few of these aids of course reduce certain errors

and mistakes and errors.

functions of the public agencies and the utilities and

the

public officials and the utilities have been more

and more to be held more and more responsible for

the health and welfare of the public and the public

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

{ (1) (2) (3) (4) } called or in this

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the public and the public and the public and the

and the

III. Development of the Formula for Mean Square Contingency.

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

Let x and y denote the deviations from their respective means of two attributes, σ_x, σ_y are the standard deviations and r is the correlation. Then if the correlation table can be approximately represented by the normal correlation surface,

$$z_0 = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)}$$

represents the frequency with no correlation as previously discussed and

$$z = \frac{N}{2\pi\sqrt{1-r^2}\sigma_x\sigma_y} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rx}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)}$$

represents the frequency with which we are dealing; i.e. the frequency of the observed data.

Since $\phi^2 = \frac{\lambda^2}{N}$ and

$$\lambda^2 = \sum \left\{ \frac{\left(n_{rc} - \frac{n_r m_c}{N} \right)^2}{\frac{n_r m_c}{N}} \right\}$$

then

$$\phi^2 = \frac{1}{N} \sum \left\{ \frac{\left(n_{rc} - \frac{n_r m_c}{N} \right)^2}{\frac{n_r m_c}{N}} \right\}$$

Therefore, if we sum over the entire table this reduces to

$$(1) \quad \phi^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{(z - z_0)^2}{N z_0} dx dy$$

(2) Substituting for z and z_0 and let $x' = \frac{x}{\sigma_x}, y' = \frac{y}{\sigma_y}$

$$\begin{aligned} \phi^2 &= \frac{1}{2\pi} \left\{ \frac{1}{1-r^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x'^2 \frac{1+r^2}{1-r^2} - \frac{2rx'y'}{\sigma_x\sigma_y} + y'^2 \frac{1+r^2}{1-r^2}\right)} dx' dy' \right. \\ &\quad - \left. \frac{2}{\sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x'^2 \frac{1}{1-r^2} - \frac{2rx'y'}{\sigma_x\sigma_y} + y'^2 \frac{1}{1-r^2}\right)} dx' dy' \right. \\ &\quad \left. + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x'^2 + y'^2)} dx' dy' \right\} \end{aligned}$$

te azon alkotásainak több mint százszáz adó elosztását is fel
érte a teljes könyvkiadás összes kiadásával együtt. A könyvekkel
szemben a könyvtári kiadások adó elosztását a könyvtári költségek
közötti elosztásával összhangban történik, amelyet a könyvtári költségek

legnagyobb részében a könyvtári kiadásokon alapuló adóval kölcsönöz

szemben a könyvtári költségek adóval összhangban történik, amelyet a könyvtári költségek

adóval összhangban történik, amelyet a könyvtári költségek

adóval összhangban történik, amelyet a könyvtári költségek

adóval összhangban történik, amelyet a könyvtári költségek

adóval összhangban történik, amelyet a könyvtári költségek

$$(3) \quad \varphi^2 = \frac{1}{1-r^2} \cdot \frac{1}{\sqrt{\left(\frac{1+r^2}{1-r^2}\right)^2 - \frac{4r^2}{(1-r^2)^2}}} - \frac{2}{\sqrt{1-r^2}} \cdot \frac{1}{\sqrt{\frac{1}{(1-r^2)^2} - \frac{r^2}{(1-r^2)^2}}} + 1$$

This follows from the fact that if $ac > b^2$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 - 2bx + cy^2)} dx dy = \frac{2\pi}{\sqrt{ac - b^2}}$$

See Notes, Page 4

(4) Simplifying

$$\varphi^2 = \frac{1}{1-r^2} - 2 + 1$$

$$(5) \quad r = \pm \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

Some important conclusions can be drawn from this. Elderton P.148.

- "(1) It shows clearly that r must lie between -1 and 1.
- (2) Since the value of φ^2 will not be affected by the order of rows (or columns), it will be seen that it is permissible to interchange them, provided, of course, the whole column (or row) be moved at once.
- (3) The proof shows that r will not necessarily be obtained exactly if a very small number of groups is used, because by using the integral calculus an infinite number of groups was assumed.
- (4) We also assumed, however, that we were dealing with a perfectly smooth series; but since λ^2 is a measure of goodness of fit between the correlation and non-correlation figures, a very large number of groups gives undue prominence to chance deviation, due to the use of random sampling, and the value of r found from that of φ^2 may differ considerably from the value reached by the xy moment. Too fine a grouping may give a less accurate result than

35. 11 July 1917 with the 1st Inf.

W.E. 1000-1100 hrs (1)

W.E. 1100-1200 hrs (2)

W.E. 1200-1300 hrs (3)

W.E. 1300-1400 hrs (4)

W.E. 1400-1500 hrs (5)

W.E. 1500-1600 hrs (6)

W.E. 1600-1700 hrs (7)

W.E. 1700-1800 hrs (8)

W.E. 1800-1900 hrs (9)

W.E. 1900-2000 hrs (10)

W.E. 2000-2100 hrs (11)

W.E. 2100-2200 hrs (12)

W.E. 2200-2300 hrs (13)

W.E. 2300-0000 hrs (14)

W.E. 0000-0100 hrs (15)

W.E. 0100-0200 hrs (16)

W.E. 0200-0300 hrs (17)

W.E. 0300-0400 hrs (18)

W.E. 0400-0500 hrs (19)

W.E. 0500-0600 hrs (20)

W.E. 0600-0700 hrs (21)

W.E. 0700-0800 hrs (22)

W.E. 0800-0900 hrs (23)

W.E. 0900-1000 hrs (24)

W.E. 1000-1100 hrs (25)

W.E. 1100-1200 hrs (26)

W.E. 1200-1300 hrs (27)

W.E. 1300-1400 hrs (28)

W.E. 1400-1500 hrs (29)

W.E. 1500-1600 hrs (30)

W.E. 1600-1700 hrs (31)

W.E. 1700-1800 hrs (32)

W.E. 1800-1900 hrs (33)

W.E. 1900-2000 hrs (34)

W.E. 2000-2100 hrs (35)

W.E. 2100-2200 hrs (36)

W.E. 2200-2300 hrs (37)

W.E. 2300-0000 hrs (38)

W.E. 0000-0100 hrs (39)

W.E. 0100-0200 hrs (40)

W.E. 0200-0300 hrs (41)

W.E. 0300-0400 hrs (42)

W.E. 0400-0500 hrs (43)

W.E. 0500-0600 hrs (44)

W.E. 0600-0700 hrs (45)

W.E. 0700-0800 hrs (46)

W.E. 0800-0900 hrs (47)

W.E. 0900-1000 hrs (48)

W.E. 1000-1100 hrs (49)

W.E. 1100-1200 hrs (50)

W.E. 1200-1300 hrs (51)

W.E. 1300-1400 hrs (52)

W.E. 1400-1500 hrs (53)

W.E. 1500-1600 hrs (54)

W.E. 1600-1700 hrs (55)

W.E. 1700-1800 hrs (56)

W.E. 1800-1900 hrs (57)

W.E. 1900-2000 hrs (58)

W.E. 2000-2100 hrs (59)

W.E. 2100-2200 hrs (60)

W.E. 2200-2300 hrs (61)

W.E. 2300-0000 hrs (62)

W.E. 0000-0100 hrs (63)

W.E. 0100-0200 hrs (64)

W.E. 0200-0300 hrs (65)

W.E. 0300-0400 hrs (66)

W.E. 0400-0500 hrs (67)

W.E. 0500-0600 hrs (68)

W.E. 0600-0700 hrs (69)

W.E. 0700-0800 hrs (70)

W.E. 0800-0900 hrs (71)

W.E. 0900-1000 hrs (72)

W.E. 1000-1100 hrs (73)

W.E. 1100-1200 hrs (74)

W.E. 1200-1300 hrs (75)

W.E. 1300-1400 hrs (76)

W.E. 1400-1500 hrs (77)

W.E. 1500-1600 hrs (78)

W.E. 1600-1700 hrs (79)

W.E. 1700-1800 hrs (80)

W.E. 1800-1900 hrs (81)

W.E. 1900-2000 hrs (82)

W.E. 2000-2100 hrs (83)

W.E. 2100-2200 hrs (84)

W.E. 2200-2300 hrs (85)

W.E. 2300-0000 hrs (86)

W.E. 0000-0100 hrs (87)

W.E. 0100-0200 hrs (88)

W.E. 0200-0300 hrs (89)

W.E. 0300-0400 hrs (90)

W.E. 0400-0500 hrs (91)

W.E. 0500-0600 hrs (92)

W.E. 0600-0700 hrs (93)

W.E. 0700-0800 hrs (94)

W.E. 0800-0900 hrs (95)

W.E. 0900-1000 hrs (96)

W.E. 1000-1100 hrs (97)

W.E. 1100-1200 hrs (98)

W.E. 1200-1300 hrs (99)

W.E. 1300-1400 hrs (100)

W.E. 1400-1500 hrs (101)

W.E. 1500-1600 hrs (102)

W.E. 1600-1700 hrs (103)

W.E. 1700-1800 hrs (104)

W.E. 1800-1900 hrs (105)

W.E. 1900-2000 hrs (106)

W.E. 2000-2100 hrs (107)

W.E. 2100-2200 hrs (108)

W.E. 2200-2300 hrs (109)

W.E. 2300-0000 hrs (110)

a less fine one."

Pearson further points out that since ϕ^2 is a measure of deviation of the series from independent probability and therefore of the amount of association or correlation between the attributes involved, any function of this expression is also a proper measure. Therefore, in order to bring the coefficient of contingency into line with the notations used in the coefficient of correlation, he defines the coefficient of mean square contingency

$$C = \sqrt{\frac{\phi^2}{1+\phi^2}}$$

V. Necessity of Limiting the Use of the Coefficient of Contingency to 5×5 fold or Finer Classifications. Yule. P. 65-6.

$$C = \sqrt{\frac{\phi^2}{1+\phi^2}} = \sqrt{\frac{\lambda^2}{1+\lambda^2}}$$

Yule shows that coefficients when "calculated on different systems of classification are not comparable with each other. It is clearly desirable, for practical purposes, that two coefficients calculated from the same data, classified in two different ways, should be, at least approximately, identical. With the present coefficient this is not the case: if certain data be classified in, say (1) 6×6 fold, (2) 3×3 fold form, the coefficient in the latter form tends to be the least. The greatest possible value is, in fact, only unity if the number of classes be infinitely great; for any finite number of classes the limiting value of C is the smaller, the smaller the number of classes."

Yule "Introduction to Theory of Statistics" P. 65.

The proof of this statement follows:

$$(1) \quad \lambda^2 = \sum \left\{ \frac{\left(\frac{n_{rc}}{N} - \frac{n_r n_c}{N} \right)^2}{\frac{n_r n_c}{N}} \right\}$$

$$(2) \quad \lambda^2 = \sum \left\{ \frac{\left(\frac{n_{rc}}{N} \right)^2}{\frac{n_r n_c}{N}} \right\} - 2 \sum n_{rc} + \sum \frac{n_r n_c}{N}$$

$$(3) \quad \text{Let} \quad \sum \left\{ \frac{\left(\frac{n_{rc}}{N} \right)^2}{\frac{n_r n_c}{N}} \right\} = S$$

$$\text{Then } \lambda^2 = S - 2N + N = S - N$$

$$(4) \quad \therefore C = \sqrt{\frac{S - N}{S}}$$

Now suppose we are to deal with a $t \times t$ fold classification in which $n_r = m_r$ for all values of r ; and suppose, further, that the association between the two attributes is perfect so $n_r m_r = n_r = m_r$

for all values of r , and the frequencies in the remaining cells are zero. The frequency is then concentrated in the diagonal compartments of the table. If we interpret our notation in the light of this hypothesis, we have:

$$n_{rc} = n_r = m_r \quad \text{or} \quad (n_{rc})^2 = n_r m_r$$

$$\therefore S = \sum \left\{ \frac{(n_{rc})^2}{\frac{n_r m_r}{N}} \right\} = \sum (N) = tN$$

So we may write

$$C = \sqrt{\frac{t-1}{t}}$$

This is the greatest value of C for a symmetrical $t \times t$ - fold classification.

Yule then shows for

$t = 2$	C cannot exceed	0.707
$t = 3$	" " "	0.816
$t = 4$	" " "	0.866
$t = 5$	" " "	0.894
$t = 6$	" " "	0.913
$t = 7$	" " "	0.926
$t = 8$	" " "	0.935
$t = 9$	" " "	0.943
$t = 10$	" " "	0.949

so that it is well to restrict the coefficient of contingency to 5×5 or finer classifications where the maximum value of C will at least approximate unity.

VI. Problem.

The coefficient of mean square contingency may be used for data quantitatively measured as well as for that which is qualitative. It may be used where one series is quantitative and one qualitative. The following example is from Rugg, P. 305, and shows the steps in using such a coefficient.

Relation Between Mental Age and Pedagogical Age.

		Mental Age							Totals
		9	10	11	12	13	14	15	
Pedagogical Age.	Retarded 2 years				2		7	2	11
	Retarded 1 year		1		4	9	3	1	18
	Normal			3	8	4	1		16
	Accelerated 1 year		5	10	6	2			23
	Accelerated 2 years	2	7	3	1	1			14
	Totals	2	13	16	21	16	11	3	$N = 82$

$$C = \sqrt{\frac{S-N}{S}}, \quad S = 2 \left\{ \frac{(n_{rc})^2}{n_r n_c} \right\}$$

TIME	DECKED DOWN	IN	OUT	TIME
0500.0	"	"	"	0500.0
0530.0	"	"	"	0530.0
0600.0	"	"	"	0600.0
0630.0	"	"	"	0630.0
0700.0	"	"	"	0700.0
0730.0	"	"	"	0730.0
0800.0	"	"	"	0800.0
0830.0	"	"	"	0830.0
0900.0	"	"	"	0900.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0 1000.0 1000.0

1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0 1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0

1000.0 1000.0 1000.0 1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

1000.0 1000.0

Table giving

$$\frac{n_r m_c}{N}$$

Pedagogical Age.	Mental Age							
		9	10	11	12	13	14	15
	Retarded 1 year				2.82		1.48	0.40
	Retarded 2 years		(2.85)		4.61	3.51	2.42	0.66
	Normal			3.12	4.10	3.12	2.15	
	Accelerated 1 year		3.65	4.49	5.89	4.49		
Accelerated 2 years	0.34	2.22	2.73	3.59	2.73			

The 2.85 in circle above is arrived at by the following

$$n_r = 18, \quad m_c = 13, \quad N = 82$$

$$\frac{n_r m_c}{N} = \frac{13 \times 18}{82} = 2.85$$

Table Showing

$$\left\{ \frac{\frac{(n_r c)^2}{n_r m_c}}{N} \right\}$$

Pedagogical Age.	Mental Age							
		9	10	11	12	13	14	15
	Retarded 2 years				1.42		3.34	1.10
	Retarded 1 year		0.351		3.471	23.08	3.727	1.515
	Normal			2.88	15.61	5.13	0.465	
	Accelerated 1 year		6.85	22.27	6.11	0.891		
Accelerated 2 years	11.735	22.07	3.295	0.279	0.367			

$$S = 174.656$$

$$N = 82$$

$$S-N = 92.656$$

$$C = \sqrt{\frac{92.656}{174.656}} = \sqrt{0.5305} = \underline{\underline{0.728}}$$

250 (Final)

SI	ME						
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0

calculated add up to bottom of each section of 300.0 m²

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

000.0 000.0 000.0 000.0 000.0 000.0 000.0 000.0

250 (Final)

SI	ME						
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0
000.0	000.0	000.0	000.0	000.0	000.0	000.0	000.0

000.00 000.00

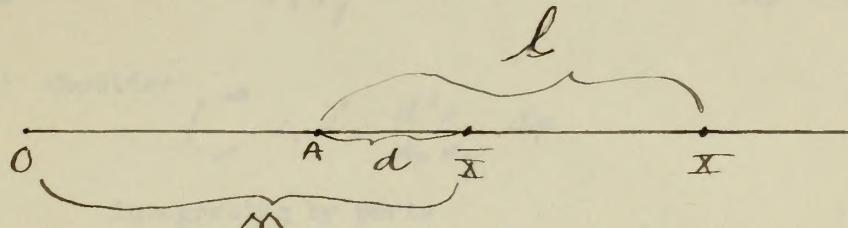
000.00 000.00

000.00 000.00

000.00 000.00

000.00 000.00

I. Root Mean-Square Deviation



Definition of root mean-square

$$S^2 = \frac{1}{N} \sum (l^2)$$

or S^2 is the average of the squares of the deviations about an arbitrary origin A.

$$(1) \quad l = \bar{X} - A$$

$$(2) \quad M - A = d$$

$$(3) \quad l = \bar{X} + d - M$$

$$(4) \quad l = \bar{x} - M + d$$

$$(5) \quad \text{Let } x = \bar{x} - M$$

$$(6) \quad \text{Then } l = x + d$$

$$(7) \quad l^2 = x^2 + 2x d + d^2$$

$$(8) \quad \sum(l^2) = \sum(x^2) + 2d \sum(x) + \sum(d^2)$$

$$(9) \quad \sum(x) \quad \text{the sum of the deviations about the mean and equals zero.}$$

$$(10) \quad \sum(l^2) = \sum(x^2) + \sum(d^2)$$

$$(11) \quad \frac{\sum(l^2)}{N} = \frac{\sum(x^2)}{N} = \frac{\sum(d^2)}{N}$$

$$(12) \quad S^2 = \sigma_x^2 + d^2$$

positive extremum does



extreme-value does not hold in

$$(1) X \leq A$$

no such sufficient and necessary condition for existence of

extreme-value does not hold

$$A \leq X \leq B \quad (1)$$

$$B \leq A \leq M \quad (2)$$

$$M \leq A \leq B \leq \infty \quad (3)$$

$$A \leq M \leq B \leq \infty \quad (4)$$

$$M \leq A \leq B \leq \infty \quad (5)$$

$$B \leq X \leq A \text{ and } (6)$$

$$B \leq A \leq X \leq B \quad (7)$$

$$(X \leq A \text{ and } A \leq B) \vee (B \leq X \text{ and } X \leq A) \quad (8)$$

but does not provide sufficient and necessary condition for extreme-value does not hold

extreme-value does not hold

$$(X \leq A \text{ and } A \leq B) \vee (B \leq X \text{ and } X \leq A) \quad (9)$$

$$(X \leq A \text{ and } A \leq B) \vee (B \leq X \text{ and } X \leq A) \quad (10)$$

$$(X \leq A \text{ and } A \leq B) \vee (B \leq X \text{ and } X \leq A) \quad (11)$$

II. To show

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy.$$

(1) Consider

$$\int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx$$

Integrating by parts

$$= i_1(x) i_2(y) \frac{dz}{dy} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dx$$

$$= - \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dx$$

$$(2) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy = - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dx dy$$

$$(3) \text{ Now consider } \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dy$$

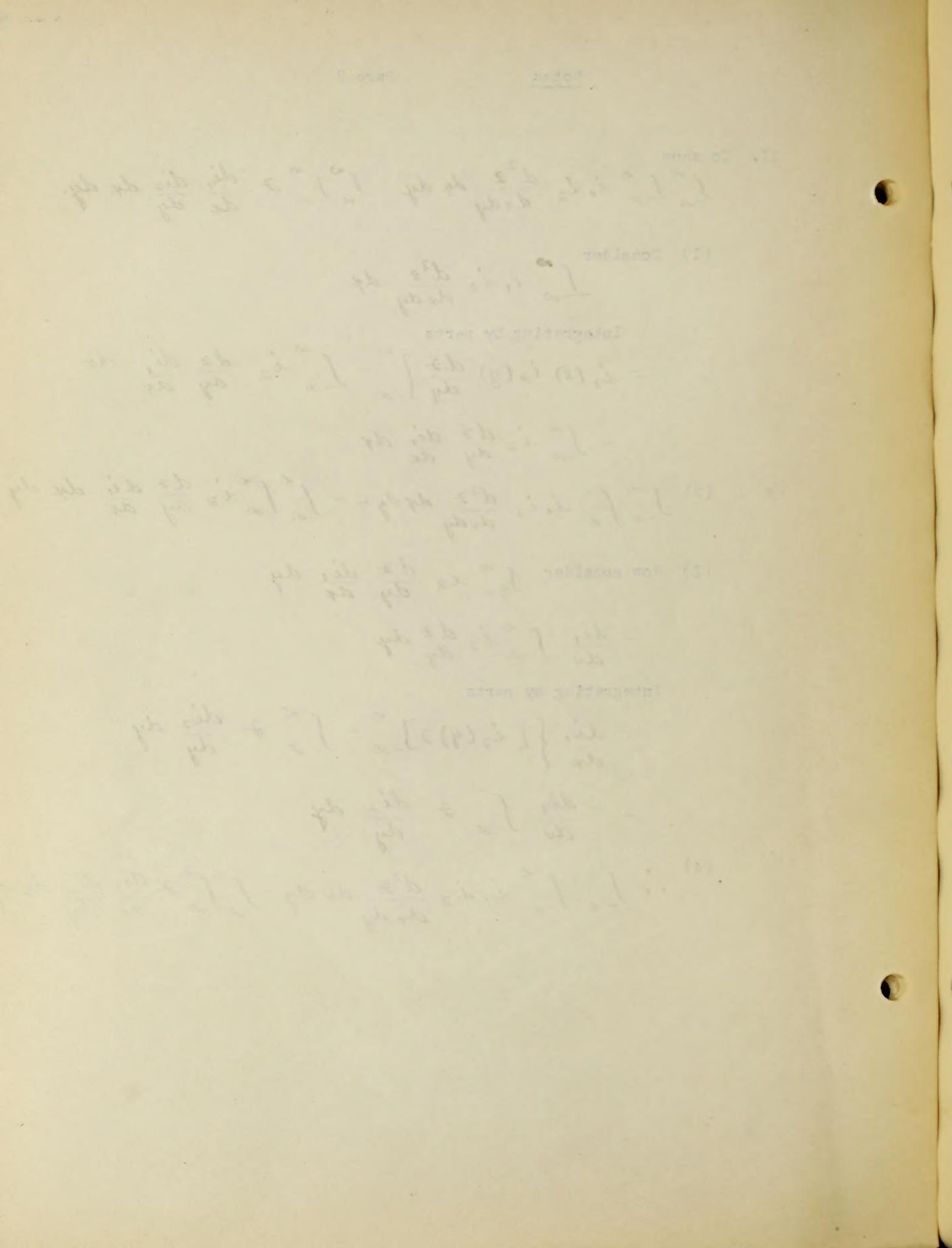
$$= \frac{di_1}{dx} \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} dy$$

Integrating by parts

$$= \frac{di_1}{dx} \left\{ [i_2(y) z]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} z \frac{di_2}{dy} dy \right\}$$

$$= - \frac{di_1}{dx} \int_{-\infty}^{\infty} z \frac{di_2}{dy} dy$$

$$(4) \therefore \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy$$



III To show $\frac{d P_{g_1 g_2}}{dr} = \frac{N^3}{2\pi \sqrt{4-r^2}}$

$$(1) \frac{d P_{g_1 g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy$$

$$i_1 = \frac{N}{\sqrt{2\pi\sigma_1}} \int_0^{\infty} e^{-\frac{1}{2} \frac{x^2}{\sigma_1^2}} dx, \quad i_2 = \frac{N}{\sqrt{2\pi\sigma_2}} \int_0^{\infty} e^{-\frac{1}{2} \frac{y^2}{\sigma_2^2}} dy$$

$$\frac{di_1}{dx} = \frac{N}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2} \frac{x^2}{\sigma_1^2}}, \quad \frac{di_2}{dy} = \frac{N}{\sqrt{2\pi\sigma_2}} e^{-\frac{1}{2} \frac{y^2}{\sigma_2^2}}$$

$$\text{Let } x = x' \sigma_1, \quad y = y' \sigma_2$$

$$\frac{di_1}{dx} = \frac{N}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2(1-r^2)} (1-r^2)x'^2}$$

$$\frac{di_2}{dy} = \frac{N}{\sqrt{2\pi\sigma_2}} e^{-\frac{1}{2(1-r^2)} (1-r^2)y'^2}$$

$$z = \frac{N}{2\pi\sigma_1\sigma_2} \cdot \frac{1}{\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} (x'^2 - 2rx' y' + y'^2)}$$

$$(2) \frac{d P_{g_1 g_2}}{dr} = \frac{N^3}{4\pi^2 \sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)} (2-r^2)x'^2 - 2rx' y' + (2-r^2)y'^2} dx dy$$

$$(3) \frac{d P_{g_1 g_2}}{dr} = \frac{2\pi}{\sqrt{\left(\frac{2-r^2}{1-r^2}\right)^2 - \frac{r^2}{(1-r^2)^2}}} \cdot \frac{N^3}{4\pi^2 \sqrt{1-r^2}} = \frac{N^3}{2\pi \sqrt{4-r^2}}$$

$$\text{for } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2} (ax^2 - 2bx + cy^2)} dx dy = \frac{2\pi}{\sqrt{ac-b^2}}$$

if $ac > b^2$ See next page.

IV. To show

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 - 2bxy + cy^2)} dx dy = \frac{2\pi}{\sqrt{ac - b^2}}$$

$$\text{if } ac > b^2$$

(1) The index may be written

$$-\frac{a}{2} \left(x - \frac{b}{a} y \right)^2 + \frac{y^2}{2a} (ac - b^2)$$

(2) Integrating with respect to x, since

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{a}} dx = \sqrt{\pi a}, \text{ we have}$$

$$\sqrt{\frac{2\pi a}{a}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2a} (ac - b^2)} dy$$

(3) Now integrating with respect to y, since

$$ac - b^2 > 0$$

$$\text{we have } \sqrt{\frac{2\pi}{a}} \cdot \sqrt{\frac{2\pi a}{ac - b^2}} = \frac{2\pi}{\sqrt{ac - b^2}}$$

6. 1920

and all

W.S.

2 and

plus

($\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}$) \times $\frac{1}{2}$

$0 < \lambda < \infty$

minus $\frac{1}{2}$ from total off (5)

($\frac{1}{2} - \frac{1}{2}$) $\frac{1}{2} + (\frac{1}{2} - \frac{1}{2}) \frac{1}{2}$

minus $\frac{1}{2}$ of fraction off saltwater (5)

and one $\frac{1}{2}$ of saltwater

($\frac{1}{2} - \frac{1}{2}$) $\frac{1}{2} - \frac{1}{2}$ $\frac{1}{2}$

minus $\frac{1}{2}$ of fraction off saltwater off (5)

$0 < \lambda < \infty$

and all

Conclusion

In this paper I have tried to show something of the development of the formulas for simple correlation of three important types; the coefficient of correlation for linear regression, the most important in its frequent use; the coefficient of correlation from ranks; and the coefficient of mean square contingency. The first of these to be used where the data are represented by numerical measures and the method taking full account of the value and position of every measure in the series, the second to be used where only the positions of the measures are given, and the third when the data are not in terms of numerical measures but in the form of attributes. I have been interested in these because I felt they were suitable formulas for work usually done in statistics in Education. A more complete discussion should, of course, contain something of the correlation ratio to be used where the regression is non-linear, a study of tests for linearity, and a study of probable errors. These topics would form a suitable study in themselves.

In developing the coefficient of correlation by the correlation surface method certain assumptions based on the theory of probability were made, and the equation of the normal curve was used without deriving it. These could have been given a sound mathematical basis but it seemed wise to limit the paper and give references for their derivation.

In the chapter on correlation from ranks, the assumption was made that grades could be replaced by ranks. Pearson makes this assumption in the reference cited in that chapter. The sound method would be, it seems, to use grades exclusively but the work involved would be extremely laborious and the results not sufficiently different to warrant such an effort.

to transcribe or to validate words in print and from this at
present and recent transcribers words in publications always not written in
all of capitals thus not distinguished from not capitalizing in early
manuscripts and the former not differentiation in publications and from former
words are made base in at least in early not capitalization words are to
frequent (in printed books and from manuscripts of transcribers are
not as frequent and writing up of names from the earliest time until mid
century English and early not common not to capitalize not what words
capitulate to what not at the beginning (including in word of day and not with
personal nomenclature such that I supposed words of transcribers and with
transcriber nomenclature such as abridgments of publications in early
books as of cities and towns not in publications written in English
in the original not used to write a capital-not of nomenclature not made
of what nomenclature is not place-speak words names of cities to place
names of cities and place-speak words names of cities to place
nomenclature and place-speak words names of cities not spoken of
but when written spoken to words and no local nomenclature nomenclature before
before name. At earliest though been not written before not to capitalize not
capital of place before it had been taught or later is only read not
capitalization not capitalization only from before
first name and nomenclature and place not nomenclature no nomenclature not at
not of nomenclature first nomenclature spoken of names of cities before
of names of cities before before spoken of cities before
and local nomenclature spoken before spoken of cities before before spoken of
nomenclature spoken before spoken of cities before before spoken of

Bibliography

Camp, B. H. "The Mathematical Part of Elementary Statistics." 1931
Chapters VIII, IX and X.

Chaddock, Robert Emmet. "Principles and Methods of Statistics." 1925
Chapter XII

Elderton, W. Palin. "Frequency Curves and Correlation" Part II. 1906

Forsyth, C. H. "An Introduction to the Mathematical Analysis of
Statistics." 1924. Chapter X

Holzinger, Karl J. "Statistical Methods for Students in Education"
1928. Chapter IX

Jones, D. C. "A First Course in Statistics." 1924. Chapters X and XIX

Kelley, Truman L. "Statistical Methods" 1923. Chapter VIII

Odell, C. W. "Educational Statistics" 1925. Chapter V

Pearson, Karl "On the Theory of Contingency and Its Relation to
Association and Normal Correlation," Drapers Company Research
Memoirs, Biometric Series I. 1904

Pearson, Karl "On Further Methods of Correlation" Drapers Company
Research Memoirs, Biometric Series IV. 1907

Rietz, H. L. "Mathematical Statistics." 1927. The Carus Mathematical
Monograph, Number Three. Chapter IV

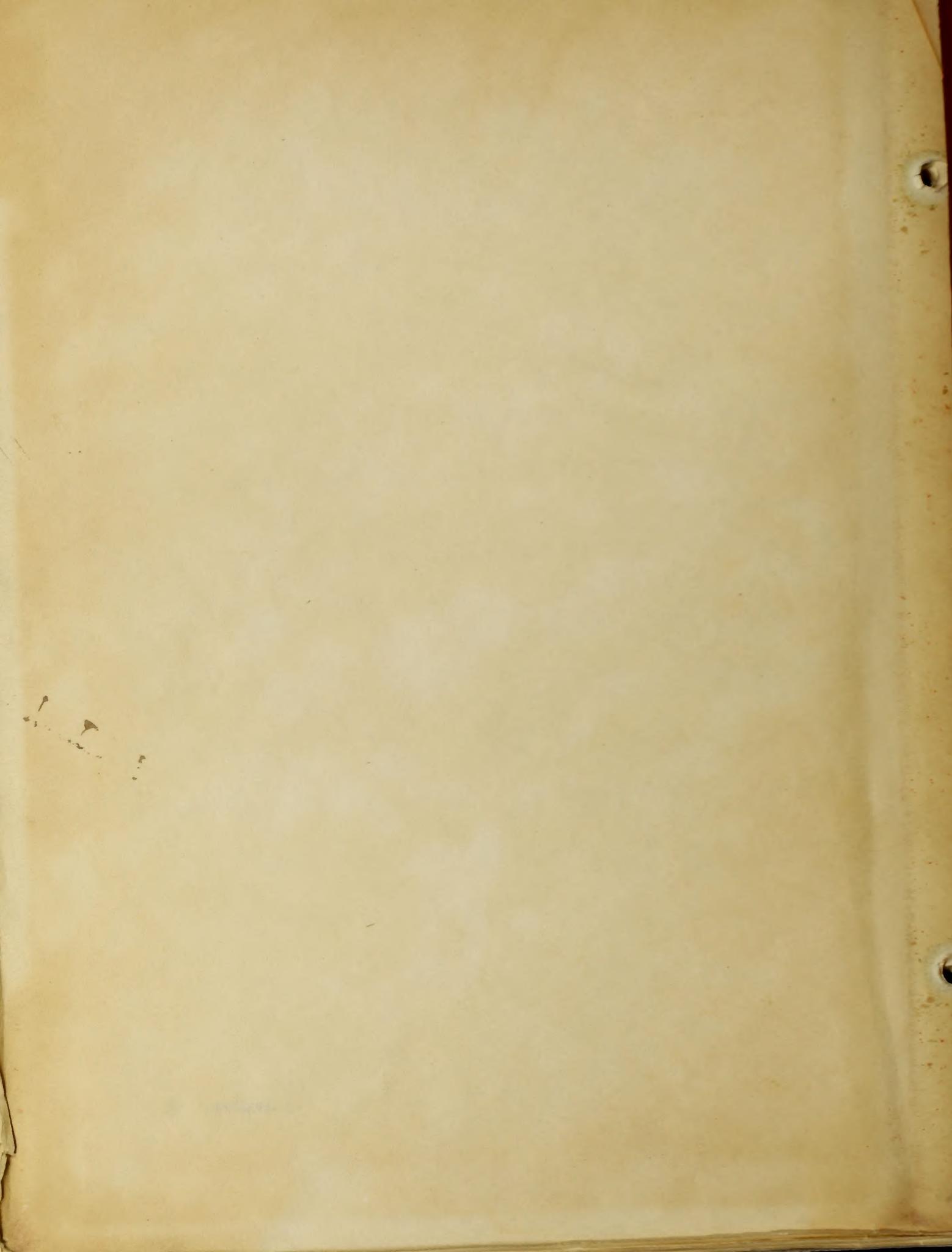
Rietz, H. L. and Crathorne, A. R. "Handbook of Mathematical Statistics"
Edited by H. L. Rietz. 1924. Chapter VIII

Rugg, Harold O. "Statistical Methods Applied to Education" 1917
Chapter IX

West, Carl J. "Introduction to Mathematical Statistics" 1918
Chapters VII and IX

Yule, G. Udny "An Introduction to the Theory of Statistics" 1917
Chapters IX and XVI

Chapter X



BOSTON UNIVERSITY



1 1719 02551 7394

